Stochastic games with the average reward

Stochastic games with the average reward

ISBN 90-9012162-5

Stochastic games with the average reward

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
prof.dr. A.C. Nieuwenhuijzen Kruseman,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op woensdag 18 november 1998 om 16.00 uur

door

János Flesch

Promotor:

Prof.dr.ir.drs. O.J. Vrieze

Copromotor:

Dr. F. Thuijsman

Beoordelingscommissie:

Prof.dr. S.H. Tijs (voorzitter)

Prof.dr. A. Hordijk (Universiteit Leiden)

Prof.dr. H.J.M. Peters

Dr. A.J. Vermeulen

Prof.dr. N. Vieille (Université Paris-Dauphine)

Stochastic games with the average reward

Proefschrift Universiteit Maastricht
Samenvatting in het Nederlands en in het Hongaars.
Trefwoorden: stochastic games, average reward, discounted reward, optimality, equilibrium.

Acknowledgements

When applying for the position as a research assistant, I only had a vague idea about what game theory entailed. Fortunately, the excellent guidance of my supervisors, Koos Vrieze and Frank Thuijsman, proved to be invaluable and I soon acquired the knowledge necessary to start my research project. Throughout the past four years we have succeeded in building a creative research team, with intensive and enthusiastic discussions. During the difficult final year, I found Koos and Frank's support and understanding immeasurable.

I wish to thank all the members of the Department of Mathematics for the pleasant working atmosphere. In particular, I would like to thank Jeroen Rutten for his contribution. It was a pleasure to share an office with him for almost four years and to converse in mathematics as well as everyday personal matters. His friendship was invaluable to me. Ranjit Bhattacharjee was also a collegue of mine, but only for a short period of time. Nevertheless, we compensated his short stay by frequent and time-consuming discussions, usually about football.

Andres Perea y Monsuwé was one of my best friends. Along with several memorable snooker matches with a final decisive black ball, I also had the pleasure of writing an article with him.

Furthermore, I would like to thank Jeroen Kuipers and Dries Vermeulen for their friendship and our discussions about the recent developements in game theory. Dries Vermeulen's skills helped me to understand some probability measures with really weird structures. Unluckily we did not have the assistance of Jeroen Kuipers, who was rather fortunate to be in Spain at the time.

Stef Tijs really amazed me with his broad knowledge. Working with such an experienced person was very constructive. I also had several interesting discussions with Nicolas Vieille and Eilon Solan, who had some excellent ideas for the open problems we studied.

Finally, I wish to express my gratitude to the following friends who contributed to my thesis: Anna Balog, Maarten Bollen, Stéfan Horváth, Veronique Janssen, Arie Koster, Mathijs Sterk, Maaike Thiadens and my sister, Ildikó Flesch.

Maastricht, September 1998

JÁNOS FLESCH

Contents

1 Introduction 1

2 The stochastic game model 5

I Zero-sum stochastic games 35

3 Simplifying optimal strategies 37

4 Improving and non-improving strategies 57

5 Markov strategies are better than stationary strategies 69

6 Almost stationary ε-equilibria 95

II General-sum stochastic games 109

7 Recursive repeated games with absorbing states 111

8 Average-discounted equilibria 123

9 More than two players 135

Concluding remarks 151

References 159

Index 165

Samenvatting (Summary in Dutch) 167

Összefoglalás (Summary in Hungarian) 169

About the author 171

Chapter 1

Introduction

This monograph is devoted to the study of stochastic games, which can be seen as decision processes with a certain number of decision makers (players). In this monograph we will always assume that there are at least two players; stochastic games with only one player are better known as Markov decision processes and the theory on such games has developed in another direction. We will now describe stochastic games with two players for the sake of simplicity; the description of stochastic games with more players is analogous. A stochastic game with two players can be given by a state space S, and related to each state $s \in S$, a bimatrix A_s in which each entry contains two real numbers (payoffs to the respective players) and a probability vector (transition vector) over the state space S. The rows and the columns of bimatrix A_s represent the available decisions (actions) in state s for player 1 and player 2 respectively. The play of the game evolves at decision moments (stages) in $\mathbb{N}$ as follows. The play starts at stage 1 in an initial state $s \in S$, where, simultaneously and independently, both players are to choose an action: player 1 has to choose a row of bimatrix A_s while player 2 has to choose a column of A_s. Each player then receives his payoff corresponding to the entry determined by these choices, and the play moves to a new state $t \in S$, according to the transition vector. In the new state t at stage 2, the players have to choose actions again, and just like before, depending on their choices, they receive the corresponding payoffs and the play moves to a new state again, and so on to infinity.

Note that the game is non-cooperative, meaning that the players are not allowed to make binding agreements. It is furthermore assumed that the players have complete information (they know the bimatrices) and have perfect recall

about the past history of the play. Consequently, when the players have to choose actions in the current state, they may take the entire past history into account.

A plan which tells a player how to make his decisions during the play is called a strategy. Instead of choosing an action with probability 1, a strategy may as well prescribe to apply a probability distribution on the set of available actions (mixed action) for the selection. The most complex strategies are the history dependent strategies, which prescribe mixed actions in the current state depending on the past history of the play. If the prescribed mixed actions in the current state only depend on the current stage then the strategy is called a Markov strategy; while if the prescribed mixed actions are fixed for each state then the strategy is called stationary.

As a result of the play, each player obtains an infinite sequence of payoffs. These sequences need to be evaluated in some manner. We will mainly consider the average reward (simply referred to as reward), which uses the long term average payoffs as an evaluation. The goal of each player in the game is simply to maximize his own reward by means of applying an effective strategy.

Zero-sum stochastic games are special stochastic games with two players who both have completely opposite interests, namely one player pays the other player and the gain of one player is the loss of the other player. We assume that player 1 is paid by player 2, hence player 1 is trying to maximize his own reward while player 2 aims to minimize player 1's reward (player 2's maximization of his own reward now coincides with player 2's minimization of player 1's reward). It is fortunate to know that there is always a certain reward which satisfies the following properties: for any $\varepsilon > 0$, player 1 has a strategy that guarantees him at least this reward (up to this ε) against any strategy of player 2, while there are available strategies for player 2 which ensure him of not needing to pay more than this reward (up to this ε) regardless of player 1's strategy. This unique reward is called the value of the game and the above strategies are called ε-optimal strategies. Clearly, the value as a reward is an acceptable outcome of the game for both players as neither of them is able to force a better reward in his favor. It is an interesting fact that 0-optimal strategies do not necessarily exist and achieving ε-optimality can often only be possible by employing history dependent strategies.

Stochastic games (not necessarily with only two players) without the requirement that the players have completely opposite interests are called general-sum stochastic games. Since, in these games, some players may as well have matching interests up to some extent, the concepts 'value' and 'optimality' are no

longer applicable. Here the usual solution concept is that of ε-equilibrium, $\varepsilon > 0$, which is a collection of strategies from the players with the property that no player can improve his own reward by more than ε if he unilaterally deviates to another strategy. Hence, for small ε, the rewards corresponding to an ε-equilibrium are an appealing solution of the stochastic game. It is known that 0-equilibria do not always exist and history dependent strategies are often indispensable for obtaining ε-equilibria. The existence of ε-equilibria in stochastic games, however, is not yet known and is the most challenging open problem these days, even though the existence problem has been answered in the affirmative for several special classes, such as in all stochastic games with only two players.

This monograph is structured as follows. Chapter 2 describes the stochastic game model in detail and provides a summary of the most important results. This is followed by several chapters on zero-sum stochastic games. Chapter 3 deals with possible simplifications of 0-optimal strategies: we show that the existence of 0-optimal strategies implies the existence of stationary ε-optimal strategies and Markov 0-optimal strategies. This means that it is unnecessary to play complex history dependent 0-optimal strategies, whenever they exist, since stationary and Markov strategies are equally effective. In chapter 4, we extend the results to possible simplifications of so-called non-improving strategies. Next, a thorough analysis on the comparison of the effectiveness of stationary strategies and Markov strategies follows in chapter 5. We present an interesting game which demonstrates the advantage of applying Markov strategies. However, we also provide several conditions under which the two classes of strategies perform equally well. Chapter 6 deals with the structure of ε-equilibria in zero-sum stochastic games (the concept of equilibria is also applicable to zero-sum stochastic games because they are special general-sum stochastic games). In the second part of this monograph, we turn our attention to general-sum stochastic games. We only consider games with two players; games with more players are treated in chapter 9. In chapter 7, we show the existence of stationary ε-equilibria under conditions imposed upon the payoff and the transition structure. Chapter 8 provides some results on the existence of equilibria when player 1 is still interested in the average reward, but player 2 uses the so-called discounted reward. Chapter 9 is devoted to stochastic games with more than two players. Finally, we make some concluding remarks about alternative evaluations similar to the average reward as well as some alternative solution concepts regarding the stochastic game model.

Chapter 2

The stochastic game model

This chapter is devoted to the description of stochastic games with two players and the discussion on the most important issues. Extensions of the model and the results to K-person stochastic games are dealt with in chapter 9.

2.1 The game and the rules

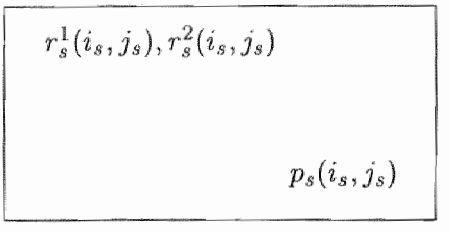
Definition 2.1.1 *A two-person stochastic game* Γ *is a tuple*

$$\left\langle S, (I_s)_{s\in S}, (J_s)_{s\in S}, (r_s^1)_{s\in S}, (r_s^2)_{s\in S}, (p_s)_{s\in S} \right\rangle ,$$

where

- S *is a nonempty and finite set, called the state space;*
- I_s *is a nonempty and finite set, called the action space for player 1 in state* $s \in S$*;*
- J_s *is a nonempty and finite set, called the action space for player 2 in state* $s \in S$*;*
- r_s^k *is a payoff function for player* $k \in \{1,2\}$ *in state* $s \in S$*, which assigns a real number* $r_s^k(i_s, j_s)$*, called payoff, to each action pair* $(i_s, j_s) \in I_s \times J_s$*;*
- p_s *is a transition map in state* $s \in S$*, which assigns a probability distribution on the state space* $p_s(i_s, j_s) = (p_s(t|i_s, j_s))_{t\in S}$*, called transition vector, to each action pair* $(i_s, j_s) \in I_s \times J_s$*.*

In the sequel, whenever we talk about stochastic games we will have two-person stochastic games in mind, unless mentioned otherwise.
In view of the above definition, a two-person stochastic game can be represented as a collection of bimatrices $\{\text{Bimatrix}(s) : s \in S\}$, where entry (i_s, j_s) of Bimatrix(s) consists of the two corresponding payoffs $r_s^k(i_s, j_s)$, $k = 1, 2$, and the corresponding transition vector $p_s(i_s, j_s)$, and is given as



When the transition vector $p_s(i_s, j_s)$ has a component $p_s(t|i_s, j_s)$ equal to 1, for some $t \in S$, then the transition vector $p_s(i_s, j_s)$ shall be abbreviated by t. Further abbreviations and notations are explained later with the help of examples 2.7.3 and 2.8.3.
In any state $s \in S$, in this bimatrix representation, the actions of player 1 are simply the rows and the actions of player 2 are simply the columns of Bimatrix(s).

The play of the game evolves at stages in $\mathbb{N}$ as follows. The play starts at stage 1 in an initial state $s^1 \in S$, where, simultaneously and independently, both players are to choose an action: player 1 chooses an $i_{s^1}^1 \in I_{s^1}$, while player 2 chooses a $j_{s^1}^1 \in J_{s^1}$. These choices induce an immediate payoff $r_{s^1}^1(i_{s^1}^1, j_{s^1}^1)$ to player 1 and an immediate payoff $r_{s^1}^2(i_{s^1}^1, j_{s^1}^1)$ to player 2. Next, the play moves to a new state according to the transition vector $p_{s^1}(i_{s^1}^1, j_{s^1}^1)$, namely a transition occurs to state $s^2 \in S$ with probability $p_{s^1}(s^2|i_{s^1}^1, j_{s^1}^1)$. At stage 2 in the new state s^2, new actions $i_{s^2}^2 \in I_{s^2}$ and $j_{s^2}^2 \in J_{s^2}$ are to be chosen by the players. Afterwards the players receive the corresponding payoffs $r_{s^2}^1(i_{s^2}^2, j_{s^2}^2)$ and $r_{s^2}^2(i_{s^2}^2, j_{s^2}^2)$, and the play moves to some state s^3 according to the transition vector $p_{s^2}(i_{s^2}^2, j_{s^2}^2)$ again, and so on to infinity.

It is assumed that the players know the structure of the game, namely the tuple in definition 2.1.1, and that they are informed about the present state and have perfect recall of the past history $(s^m, i_{s^m}^m, j_{s^m}^m)_{m=1}^{n-1}$ at any stage n of the play.

It should be noticed that, depending on the initial state, the stochastic game situation is different. However, it often appears useful to treat these games simultaneously.

2.2 Strategies

Definition 2.2.1

(a) *A sequence* $h^n = (s^m, i^m_{s^m}, j^m_{s^m})^n_{m=1}$, *with* $n \in \mathbb{N}$, *where* $s^m \in S$, $i^m_{s^m} \in I_{s^m}$, $j^m_{s^m} \in J_{s^m}$ *for all* $m = 1, \ldots, n$, *is called a history up to stage* n. *Here state* s^1 *is called the initial state of history* h^n. *The initial history up to stage* 0 *is the empty sequence* $h^0 := ()$. *Let* H^n *denote the set of histories up to stage* n, *and let* $H^0 := \{h^0\}$. *Let* $H := \cup^\infty_{n=0} H^n$ *denote the set of finite histories.*

(b) *An infinite history is a sequence* $h^\infty = (s^n, i^n_{s^n}, j^n_{s^n})_{n\in\mathbb{N}}$ *where* $s^n \in S$, $i^n_{s^n} \in I_{s^n}$, $j^n_{s^n} \in J_{s^n}$ *for all* $n \in \mathbb{N}$. *Here state* s^1 *is called the initial state of history* h^∞. *The set of infinite histories is denoted by* H^∞.

(c) *For* $s \in S$, *let* H^n_s *denote the set of histories up to stage* n *with initial state* s *for all* $n \in \mathbb{N} \cup \{0\}$, *let* H_s *denote the set of finite histories with initial state* s, *and let* H^∞_s *denote the set of infinite histories with initial state* s.

Whenever it is convenient, we will only consider histories

$$h^n = (s^m, i^m_{s^m}, j^m_{s^m})^n_{m=1}, \qquad n \geq 2,$$

which are consistent with the transition structure of the stochastic game, namely

$$p_{s^m}(s^{m+1} | i^m_{s^m}, j^m_{s^m}) > 0 \quad \forall m = 1, \ldots, n-1.$$

When playing a stochastic game, the players may randomize over their actions, namely instead of choosing an action with probability 1, they may use a probability distribution on the action spaces. Such probability distributions are called mixed actions.

Definition 2.2.2 *A mixed action* x_s *for player 1 in state* $s \in S$ *is a probability distribution on the set* I_s. *A mixed action* y_s *for player 2 in state* $s \in S$ *is a probability distribution on the set* J_s. *The respective sets of mixed actions in state* s *are denoted by* X_s *and* Y_s.

Note that the sets X_s and Y_s, $s \in S$, are nonempty polytopes. Moreover, any action $i_s \in I_s$, in any state $s \in S$, can be naturally identified with the mixed action in X_s which puts probability 1 on i_s. Actions in J_s can be similarly identified with mixed actions in Y_s, in any state $s \in S$. By these identifications, I_s and J_s become the respective sets of extreme points of X_s and Y_s.
Next we define the most important classes of strategies.

Definition 2.2.3

(a) *A (history dependent) strategy for player 1 is a map π assigning a mixed action $\pi_s(h) \in X_s$ in any present state $s \in S$ for any past history $h \in H$. A (history dependent) strategy for player 2 is a map σ assigning a mixed action $\sigma_s(h) \in Y_s$ in any present state $s \in S$ for any past history $h \in H$. For simplicity, let $\pi_s := \pi_s(h^0)$ and $\sigma_s := \sigma_s(h^0)$ for all $s \in S$. The respective sets of history dependent strategies are denoted by Π and Σ.*

(b) *A strategy π for player 1 is called a Markov strategy if $\pi_s(h^n) = \pi_s(\bar{h}^n)$ in any present state $s \in S$ for any past histories $h^n, \bar{h}^n \in H^n$, $n \in \mathbb{N}$. A strategy σ for player 2 is called a Markov strategy if $\sigma_s(h^n) = \sigma_s(\bar{h}^n)$ in any present state $s \in S$ for any past histories $h^n, \bar{h}^n \in H^n$, $n \in \mathbb{N}$. We use the notations f and g for Markov strategies of the players, while $f_s(n)$ and $g_s(n)$ for the unique mixed actions that are prescribed by the strategies f and g in present state s at stage n. The respective sets of Markov strategies are $F := \times_{s\in S, n\in\mathbb{N}} X_s$ and $G := \times_{s\in S, n\in\mathbb{N}} Y_s$.*

(c) *A strategy π for player 1 is called a stationary strategy if $\pi_s(h^n) = \pi_s(h^m)$ in any present state $s \in S$ for any past histories $h^n \in H^n$, $h^m \in H^m$, $n, m \in \mathbb{N}$. A strategy σ for player 2 is called a stationary strategy if $\sigma_s(h^n) = \sigma_s(h^m)$ in any present state $s \in S$ for any past histories $h^n \in H^n$, $h^m \in H^m$, $n, m \in \mathbb{N}$. We use the notations x and y for stationary strategies of the players, while x_s and y_s for the unique mixed actions that are prescribed by the strategies x and y in present state s. The respective sets of stationary strategies are $X := \times_{s\in S} X_s$ and $Y := \times_{s\in S} Y_s$.*

The intuition behind the definitions is the following. History dependent strategies are the most general strategies. When a player uses a history dependent strategy, he takes the entire past history into account when choosing his action in the present state. Markov strategies, however, have a substantially easier structure than history dependent strategies, since Markov strategies only consider the present state and present stage when prescribing a mixed action.

A fundamental role in the analysis of stochastic games will be played by stationary strategies, where the prescribed mixed actions only depend on the present state.
We wish to distinguish strategies which do not use randomization. These strategies always prescribe one action to be used with probability 1.

Definition 2.2.4 *A strategy $\pi \in \Pi$ for player 1 is called pure if $\pi_s(h) \in I_s$ in any present state $s \in S$ for any past history $h \in H$. For player 2, pure strategies are defined analogously. Let Π^p and Σ^p denote the respective spaces of pure (history dependent) strategies, F^p and G^p the respective spaces of pure Markov strategies, and I and J the respective spaces of pure stationary strategies. Pure stationary strategies are denoted by i for player 1 and by j for player 2.*

In fact, $I = \times_{s\in S} I_s$ and $J = \times_{s\in S} J_s$, and they are the respective sets of extreme points of the polytopes X and Y.
Finally, we define what we mean by a strategy conditional on a past history.

Definition 2.2.5 *Let $\pi \in \Pi$ and $h \in H$. The strategy π conditional on the history h, denoted by $\pi[h]$, is the strategy which prescribes a mixed action $\pi_s[h](\bar{h})$ in any present state $s \in S$ for any history $\bar{h} \in H$ as if h had happened before $\bar{h}$, namely $\pi_s[h](\bar{h}) = \pi_s(h \oplus \bar{h})$, where $h \oplus \bar{h}$ is the history consisting of h concatenated with $\bar{h}$. The definition is analogous for strategies of player 2.*

Notice that any strategy conditional on the initial history h^0 is simply itself.

2.3 Probability measures on the set of histories

Take an initial state $s \in S$ and a pair of (history dependent) strategies (π, σ). In this section, we will consider probability measures that are induced by the triple (s, π, σ) on sets of histories. The measure theoretic concepts that we will use can be found in elementary books on measure theory.
For $n \in \mathbb{N}$, consider the finite set H_s^n of histories up to stage n with initial state s. Let $\mathcal{M}_s^n$ denote the set which consists of all the subsets of H_s^n. Then the pair $(H_s^n, \mathcal{M}_s^n)$ is a measurable space, on which the triple (s, π, σ) induces a probability measure $\mathcal{P}_{s\pi\sigma}^n$ in a natural way. So, if $U \in \mathcal{M}_s^n$ then $\mathcal{P}_{s\pi\sigma}^n(U)$ gives the probability that the history up to stage n will belong to U. We have thus obtained a probability measure space $(H_s^n, \mathcal{M}_s^n, \mathcal{P}_{s\pi\sigma}^n)$ which describes the play up to stage n in a probabilistic manner. For the sake of completeness, we

also consider the probability measure space $(H_s^0, \mathcal{M}_s^0, \mathcal{P}_{s\pi\sigma}^0)$ where $H_s^0 = \{h^0\}$, $\mathcal{M}_s^0 = \{\emptyset, H_s^0\}$, $\mathcal{P}_{s\pi\sigma}^0(h^0) = 1$.
We will now define a probability measure space in order to be able to describe the infinite play in a probabilistic way, with respect to (s, π, σ). This probability measure space is generated by the family of probability measure spaces $(H_s^n, \mathcal{M}_s^n, \mathcal{P}_{s\pi\sigma}^n)$, $n \in \mathbb{N} \cup \{0\}$, as follows. We may naturally identify each history $h^n \in H_s^n$ with the set $H_s^\infty[h^n]$ of infinite histories which coincide with h^n up to stage n. Similarly, we may identify each $U \subset H_s^n$ with the set

$$H_s^\infty[U] := \cup_{h^n \in U} H_s^\infty[h^n];$$

if $U = \emptyset$ then $H_s^\infty[U] := \emptyset$. On basis of these identifications, we let

$$\mathcal{M}_s^\infty := \cup_{n \in \mathbb{N} \cup \{0\}} \mathcal{M}_s^n.$$

In fact, one can check that $\mathcal{M}_s^\infty$ is an algebra of subsets of H_s^∞. Let $\mathcal{S}(\mathcal{M}_s^\infty)$ denote the sigma-algebra generated by the algebra $\mathcal{M}_s^\infty$. It is known (cf. Kolmogorov [1933]; or Dudley [1989], theorem 3.1.1 and theorem 3.1.10) that there is a unique probability measure $\mathcal{P}_{s\pi\sigma}$ on the measurable space $(H_s^\infty, \mathcal{S}(\mathcal{M}_s^\infty))$ with the consistency property that

$$\mathcal{P}_{s\pi\sigma}(U) = \mathcal{P}_{s\pi\sigma}^n(U) \qquad \forall U \in \mathcal{M}_s^n, \ \forall n \in \mathbb{N} \cup \{0\}.$$

So we have obtained a probability measure space $(H_s^\infty, \mathcal{S}(\mathcal{M}_s^\infty), \mathcal{P}_{s\pi\sigma})$ which is consistent with the family of probability measure spaces $(H_s^n, \mathcal{M}_s^n, \mathcal{P}_{s\pi\sigma}^n)$, $n \in \mathbb{N} \cup \{0\}$, in the sense that the probability measure $\mathcal{P}_{s\pi\sigma}$ coincides with $\mathcal{P}_{s\pi\sigma}^n$ on the set $\mathcal{M}_s^n$.
From this point on, we will only deal with sets of histories that belong to $\mathcal{S}(\mathcal{M}_s^\infty)$, and whenever we talk about random variables we will have random variables with respect to the space $(H_s^\infty, \mathcal{S}(\mathcal{M}_s^\infty), \mathcal{P}_{s\pi\sigma})$ in mind. We will use the notation $\mathcal{E}_{s\pi\sigma}$ for the expectation taken with respect to $\mathcal{P}_{s\pi\sigma}$.

2.4 Stochastic processes on the set of states

A pair of history dependent strategies together with an initial state induce a stochastic process on the state space S, where the transitions are history dependent. In the case of Markov strategies, this stochastic process reduces to a nonhomogeneous Markov chain, while this Markov chain is even homogeneous when stationary strategies are used by the players.
A state is called absorbing, if the probability of leaving the state is zero for any available pair of actions; otherwise the state is called non-absorbing.

So absorbing states cannot be left with respect to any stochastic process on the state space, induced by any strategy pair. When the stochastic process enters an absorbing state we speak of absorption.
Take now a pair of stationary strategies $(x, y) \in X \times Y$. As mentioned above, we obtain a homogeneous Markov chain with respect to (x, y), without specifying the initial state now. The transition matrix for this Markov chain is denoted by $P(x, y)$. The matrix $P(x, y)$ is a stochastic matrix and entry (s, t) of $P(x, y)$, where $s, t \in S$, is given as

$$p_s(t|x_s, y_s) := \sum_{i_s \in I_s} \sum_{j_s \in J_s} x_s(i_s)\, y_s(j_s) \cdot p_s(t|i_s, j_s),$$

which equals the probability of a transition from state s to state t, if the players use the mixed actions x_s and y_s in state s. With respect to the Markov chain corresponding to (x, y), or equivalently with respect to $P(x, y)$, we can classify the states in the usual way (cf. Kemeny & Snell [1960], section 2.4). A state s is called transient if it has the property that, if the process starts in s, then visiting state s infinitely often has probability zero; otherwise the state is called recurrent. Note that a recurrent state s has the property that, if the process starts in state s, then visiting state s infinitely often has probability 1. A set of states $A \subset S$ is called closed if it has the property that, if the process starts in any state $s \in A$, then ever leaving A has probability zero. (The above probabilities are obviously taken with regard to the probability measure $\mathcal{P}_{sxy}$ on the set of infinite histories). A minimal closed set of states is called an ergodic set. It is known that ergodic sets form a disjoint partition of the set of recurrent states.
With respect to the Markov chain induced by (x, y), the n-stage transition probabilities are clearly given by the matrix $P^n(x, y)$. For completeness, we let $P^0(x, y) := I$, where I is the identity matrix of size $|S| \times |S|$. We use the notation $P^n(x, y)(s, t)$ for entry (s, t) of $P^n(x, y)$.
We define a stochastic matrix

$$Q(x, y) := \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} P^{n-1}(x, y);$$

here the limit is known to exist (cf. Doob [1953], theorem 2.1, page 175). Entry (s, t) of $Q(x, y)$ is denoted by $q_s(t|x, y)$ and expresses the expected average number of visits to state t if the stationary strategy pair (x, y) is used and the initial state is state s. Obviously, if $t \in S$ is a transient state then $q_s(t|x, y) = 0$ for all $s \in S$; while if $s, t \in S$ are recurrent then $q_s(t|x, y) > 0$

if and only if s and t belong to the same ergodic set. If $E \subset S$ is an ergodic set then the probability distribution $(q_s(t|x,y))_{t\in E}$ is the same for all $s \in E$, so the s-th row and the t-th row of $Q(x,y)$ are equal if s and t belong to the same ergodic set E. In fact, $(q_s(t|x,y))_{t\in E}$, for any $s \in E$, is the unique stationary distribution of the Markov chain corresponding to the ergodic set E if the players use (x,y).
The matrix $Q(x,y)$ has the property

$$P(x,y) \cdot Q(x,y) = Q(x,y) \cdot P(x,y) = Q(x,y), \tag{2.1}$$

which follows from the definitions. Using equations (2.1) inductively, we obtain for all $n \in \mathbb{N}$ that

$$P^n(x,y) \cdot Q(x,y) = Q(x,y) \cdot P^n(x,y) = Q(x,y).$$

Hence by the definition of $Q(x,y)$ we also have

$$Q(x,y) \cdot Q(x,y) = Q(x,y). \tag{2.2}$$

2.5 The average reward

As we have already discussed, during the play the players receive infinite sequences of payoffs. These sequences must be evaluated in some manner. We mainly deal with the so-called average reward for an evaluation of these sequences.
The average reward was introduced by Gillette [1957] and is defined as follows.

Definition 2.5.1 *The average reward with respect to a strategy pair $(\pi,\sigma) \in \Pi \times \Sigma$ and initial state $s \in S$ is defined for player $k \in \{1,2\}$ as*

$$\gamma_s^k(\pi,\sigma) := \liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathcal{E}_{s\pi\sigma}\left(R_n^k\right) = \liminf_{N\to\infty} \mathcal{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N} R_n^k\right),$$

where R_n^k denotes the random variable for the payoff for player k at stage n. We also use the vector notations

$$\gamma^k(\pi,\sigma) := \left(\gamma_s^k(\pi,\sigma)\right)_{s\in S}, \qquad \gamma_s(\pi,\sigma) := \left(\gamma_s^k(\pi,\sigma)\right)_{k=1,2}.$$

The average reward uses the long term expected average payoff for the evaluation of the infinite sequences of payoffs. As the limit does not necessarily exist, we need to take a limit point of the sequences. In the above definition

we chose the limit inferior, even though all the further results remain valid with respect to the limit superior as well. The concluding remarks of this monograph provide more details on these issues.
This monograph mainly deals with the average reward. So whenever we talk about rewards in the sequel we will have the average reward in mind, unless mentioned otherwise.

2.6 The discounted reward

In the literature of stochastic game theory, the discounted reward was the first mentioned reward in Shapley [1953], and since then it has been one of the most widely used ones.

Definition 2.6.1 *Let $\beta \in (0,1)$. The β-discounted reward with respect to a strategy pair (π,σ) and initial state $s \in S$ is defined for player $k \in \{1,2\}$ as*

$$\gamma_{\beta s}^{k}(\pi,\sigma) := (1-\beta) \cdot \sum_{n=1}^{\infty} \beta^{n-1} \cdot \mathcal{E}_{s\pi\sigma}\left(R_n^k\right),$$

where R_n^k denotes the random variable for the payoff for player k at stage n. We also use the vector notations

$$\gamma_{\beta}^{k}(\pi,\sigma) := \left(\gamma_{\beta s}^{k}(\pi,\sigma)\right)_{s\in S}, \qquad \gamma_{\beta s}(\pi,\sigma) := \left(\gamma_{\beta s}^{k}(\pi,\sigma)\right)_{k=1,2}.$$

The idea of the β-discounted reward is that the payoff at stage n has to be discounted $n-1$ times, as it is received $n-1$ stages later than the first payoff at stage 1. In economic applications, this discount factor β reflects an interest rate $(1-\beta)/\beta$. The factor $(1-\beta)$ is just a normalizing factor so that the discounted reward becomes c if all the payoffs equal the same constant c.
The discounted reward itself is not only used in economic applications, but also provides one of the most frequently used tools for the analysis of the average reward.

2.7 The rewards for stationary strategies

A pair of stationary strategies induces a homogeneous Markov chain as discussed in section 2.4. Due to the simple structure of such stochastic processes, the average reward and the discounted reward can be easier calculated for stationary strategies.

When using a pair of stationary strategies $(x, y) \in X \times Y$, the prescribed mixed actions only depend on the present state, therefore whenever a state $s \in S$ is visited, the expected payoff for player k is

$$r_s^k(x_s, y_s) := \sum_{i_s \in I_s} \sum_{j_s \in J_s} x_s(i_s)\, y_s(j_s) \cdot r_s^k(i_s, j_s)$$

and the expected transition vector is exactly the s-th row of the stochastic matrix $P(x, y)$.
We also use the vector notations

$$r^k(x, y) := (r_s^k(x_s, y_s))_{s \in S}, \qquad r_s(x_s, y_s) := (r_s^k(x_s, y_s))_{k=1,2}.$$

Lemma 2.7.1 *Let $k \in \{1, 2\}$. Take a stationary strategy pair $(x, y) \in X \times Y$. Then*

(a) $\gamma^k(x, y) = Q(x, y) \cdot r^k(x, y)$;

(b) $\gamma^k(x, y) = P(x, y) \cdot \gamma^k(x, y)$;

(c) $\gamma^k(x, y) = Q(x, y) \cdot \gamma^k(x, y)$;

(d) $\gamma_s^k(x, y) = \gamma_t^k(x, y)$ *if s and t belong to the same ergodic set for (x, y).*

Proof.
(a) Let R_n^k denote the random variable for the payoff for player k at stage n. Then for all $s \in S$ we have

$$\mathcal{E}_{sxy}\left(R_n^k\right) = \sum_{t \in S} P^{n-1}(x, y)(s, t) \cdot r_t^k(x_t, y_t).$$

Using definition 2.5.1, for all $s \in S$

$$\begin{aligned} \gamma_s^k(x, y) &= \liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mathcal{E}_{sxy}\left(R_n^k\right) \\ &= \liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{t \in S} P^{n-1}(x, y)(s, t) \cdot r_t^k(x_t, y_t) \\ &= \sum_{t \in S} q_s(t|x, y) \cdot r_t^k(x_t, y_t). \end{aligned}$$

Therefore

$$\gamma^k(x, y) = Q(x, y) \cdot r^k(x, y),$$

which completes the proof of (a).

(b) It immediately follows from (a) by using (2.1).

(c) Using (b) inductively, we have for all $n \in \mathbb{N}$ that

$$\gamma^k(x,y) = P^n(x,y) \cdot \gamma^k(x,y).$$

Now by the definition of $Q(x,y)$ we obtain (c).

(d) It is a consequence of (c) using the fact that if s and t belong to the same ergodic set for (x,y), then the s-th row and the t-th row of $Q(x,y)$ are equal. □

Next we discuss an important continuity property of the average reward on the spaces of stationary strategies.

Lemma 2.7.2 *Let $k \in \{1,2\}$. Let (x^n, y^n), $n \in \mathbb{N}$, be a sequence in $X \times Y$ converging to some (x,y) in $X \times Y$. Suppose that, for all $n \in \mathbb{N}$, the ergodic sets with respect to (x^n, y^n) coincide with the ergodic sets with respect to (x,y). Then $\gamma^k(x^n, y^n)$ has a limit as n tends to infinity and*

$$\gamma^k(x,y) = \lim_{n\to\infty} \gamma^k(x^n, y^n).$$

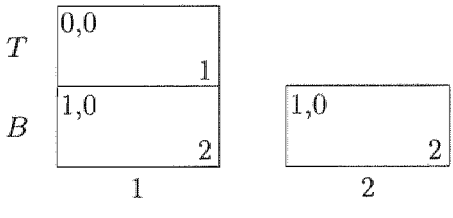
Proof. It is shown in Schweitzer [1968] (theorem 5) that

$$Q(x,y) = \lim_{n\to\infty} Q(x^n, y^n)$$

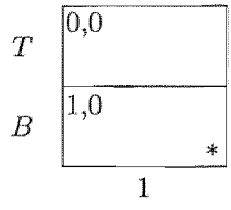
holds for such a sequence (x^n, y^n), $n \in \mathbb{N}$. Hence lemma 2.7.1-(a) implies the statement. □

Note that, in general, the average reward is not continuous on the spaces of stationary strategies, as illustrated by the next example.

Example 2.7.3



This is a stochastic game with two states. The actions of player 1 in state 1 are denoted by T (standing for top) and B (standing for bottom). Notice that state 2 is absorbing and both players have only one action in state 2, hence state 2 is not an interesting initial state and strategies only need to be defined for state 1. So assume that the initial state is state 1. By suppressing the absorbing state and the transition for action T which keeps the play in the same state with probability 1, and by denoting the absorption for action B by a $*$, we may represent the game as follows:



Such a shorter representation shall often be used later. Let y denote player 2's only strategy. Now we have $\gamma_1^1(x,y) = 1$ for all $x \in \{(u, 1-u) | \, u \in [0,1)\} \subset X$, however $\gamma_1^1((1,0),y) = 0$, which demonstrates that the average reward is not continuous on $X \times Y$. In fact, state 1 is transient with respect to (x,y) for all $x \in \{(u, 1-u) | \, u \in [0,1)\}$, while it becomes recurrent for $((1,0),y)$. This clarifies the necessity of the condition on the ergodic structures in lemma 2.7.2 above. $\lhd$

Lemma 2.7.4 *Let $k \in \{1,2\}$. Let $(x,y) \in X \times Y$ and $\beta \in (0,1)$. Then*

$$\gamma_\beta^k(x,y) = (1-\beta) \cdot \sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x,y) \cdot r^k(x,y).$$

Let I denote the identity matrix of size $|S| \times |S|$. Then the inverse of the matrix $(I - \beta \cdot P(x,y))$ exists, and

$$\gamma_\beta^k(x,y) = (1-\beta) \cdot (I - \beta \cdot P(x,y))^{-1} \cdot r^k(x,y).$$

Proof. Since

$$(I - \beta \cdot P(x,y)) \cdot \left(\sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x,y) \right) = I$$

$$\left(\sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x,y) \right) \cdot (I - \beta \cdot P(x,y)) = I,$$

we may conclude that the matrix $(I - \beta \cdot P(x,y))$ has an inverse and

$$(I - \beta \cdot P(x,y))^{-1} = \sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x,y).$$

Let R_n^k denote the random variable for the payoff for player k at stage n. As

$$\mathcal{E}_{sxy}\left(R_n^k\right) = \sum_{t \in S} P^{n-1}(x,y)(s,t) \cdot r_t^k(x_t, y_t) \quad \forall s \in S,$$

it follows from definition 2.6.1 that for all $s \in S$

$$\gamma_{\beta s}^k(x,y) = (1-\beta) \cdot \sum_{n=1}^{\infty} \beta^{n-1} \cdot \sum_{t \in S} P^{n-1}(x,y)(s,t) \cdot r_t^k(x_t, y_t).$$

Therefore

$$\begin{aligned} \gamma_\beta^k(x,y) &= (1-\beta) \cdot \sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x,y) \cdot r^k(x,y) \\ &= (1-\beta) \cdot (I - \beta \cdot P(x,y))^{-1} \cdot r^k(x,y), \end{aligned}$$

which completes the proof. □

The following continuity property of the discounted reward makes the analysis of discounted games substantially easier.

Lemma 2.7.5 *Let $k \in \{1,2\}$. The function $\gamma_\beta^k(\cdot,\cdot)$ is continuous on $X \times Y$ for any $\beta \in (0,1)$.*

Proof. It follows from the second part of lemma 2.7.4, since each factor is continuous on $X \times Y$. □

There is a strong relation between the average reward and the discounted rewards for stationary strategies. This is stated in the next lemma.

Lemma 2.7.6 *Let $k \in \{1,2\}$. Let (x^n, y^n), $n \in \mathbb{N}$, be a sequence in $X \times Y$ converging to some (x,y) in $X \times Y$. Suppose that, for all $n \in \mathbb{N}$, the ergodic sets with respect to (x^n, y^n) coincide with the ergodic sets with respect to (x,y). Let β_n, $n \in \mathbb{N}$, be a sequence of discount factors in $(0,1)$ converging to 1. Then $\gamma_{\beta_n}^k(x^n, y^n)$ has a limit as n tends to infinity and*

$$\gamma^k(x,y) = \lim_{n \to \infty} \gamma_{\beta_n}^k(x^n, y^n).$$

In particular, for any $(x, y) \in X \times Y$ *we have*

$$\gamma^k(x, y) = \lim_{\beta \uparrow 1} \gamma^k_\beta(x, y).$$

Proof. First, let E be an ergodic set for (x, y), or equivalently for (x^n, y^n) for all $n \in \mathbb{N}$. Then, by lemmas 2.2.5 and 2.2.6 in Thuijsman [1992], we obtain

$$\gamma^k_s(x, y) = \lim_{n \to \infty} \gamma^k_{\beta_n s}(x^n, y^n) \qquad \forall s \in E. \tag{2.3}$$

So it remains to show

$$\gamma^k_s(x, y) = \lim_{n \to \infty} \gamma^k_{\beta_n s}(x^n, y^n), \tag{2.4}$$

if the initial state s is transient for (x, y). Take an arbitrary transient initial state s. We will prove (2.4) by showing that, for any $\varepsilon > 0$, if n is sufficiently large then

$$\left| \gamma^k_s(x, y) - \gamma^k_{\beta_n s}(x^n, y^n) \right| \leq \varepsilon. \tag{2.5}$$

Let $\varepsilon > 0$. By using lemma 2.7.1-(b) inductively, we have

$$\gamma^k_s(x, y) = \sum_{t \in S} P^N(x, y)(s, t) \cdot \gamma^k_t(x, y) \qquad \forall N \in \mathbb{N}. \tag{2.6}$$

Let $\mathcal{R}$ denote the set of recurrent states for (x, y). Since s is transient for (x, y), the probability that, with respect to (x, y), the play enters an ergodic set before stage N converges to 1, as N tends to infinity. Hence we can choose a large N such that

$$\sum_{t \in S \setminus \mathcal{R}} P^N(x, y)(s, t) \leq \frac{\varepsilon}{4M}, \tag{2.7}$$

where $M > 0$ and

$$M \geq |r^k_z(i_z, j_z)| \qquad \forall i_z \in I_z,\ \forall j_z \in J_z,\ \forall z \in S.$$

Therefore (2.6) and (2.7) imply

$$\left| \gamma^k_s(x, y) - \sum_{t \in \mathcal{R}} P^N(x, y)(s, t) \cdot \gamma^k_t(x, y) \right| \leq$$

$$\leq \left| \sum_{t \in S \setminus \mathcal{R}} P^N(x, y)(s, t) \cdot \gamma^k_t(x, y) \right|$$

$$\leq \sum_{t\in S\setminus\mathcal{R}} P^N(x,y)(s,t)\cdot\left|\gamma_t^k(x,y)\right|$$

$$\leq \sum_{t\in S\setminus\mathcal{R}} P^N(x,y)(s,t)\cdot M$$

$$\leq \frac{\varepsilon}{4}. \tag{2.8}$$

By continuity and (2.3), if n is sufficiently large then

$$\left|\sum_{t\in\mathcal{R}} P^N(x,y)(s,t)\cdot\gamma_t^k(x,y) - \right.$$

$$\left. -\sum_{t\in\mathcal{R}} P^N(x^n,y^n)(s,t)\cdot\gamma_{\beta_n t}^k(x^n,y^n)\right| \leq \frac{\varepsilon}{4}. \tag{2.9}$$

By using (2.7), we also have

$$\left|\sum_{t\in\mathcal{R}} P^N(x^n,y^n)(s,t)\cdot\gamma_{\beta_n t}^k(x^n,y^n) - \sum_{t\in S} P^N(x^n,y^n)(s,t)\cdot\gamma_{\beta_n t}^k(x^n,y^n)\right|$$

$$\leq \left|\sum_{t\in S\setminus\mathcal{R}} P^N(x^n,y^n)(s,t)\cdot\gamma_{\beta_n t}^k(x^n,y^n)\right|$$

$$\leq \sum_{t\in S\setminus\mathcal{R}} P^N(x^n,y^n)(s,t)\cdot M$$

$$\leq \frac{\varepsilon}{4}. \tag{2.10}$$

In view of lemma 2.7.4, for all $n\in\mathbb{N}$

$$\gamma_{\beta_n}^k(x^n,y^n) - P^N(x^n,y^n)\cdot\gamma_{\beta_n}^k(x^n,y^n) =$$

$$= (1-\beta_n)\cdot\sum_{m=1}^{N}(\beta_n)^{m-1}\cdot P^{m-1}(x^n,y^n)\cdot r^k(x^n,y^n).$$

Therefore for large n, when β_n is large, we obtain

$$\left|\sum_{t\in S} P^N(x^n,y^n)(s,t)\cdot\gamma^k_{\beta_n t}(x^n,y^n)-\gamma^k_{\beta_n s}(x^n,y^n)\right| =$$

$$= \left|(1-\beta_n)\cdot\sum_{m=1}^{N}(\beta_n)^{m-1}\sum_{t\in S}P^{m-1}(x^n,y^n)(s,t)\cdot r^k_t(x^n_t,y^n_t)\right|$$

$$\leq \frac{\varepsilon}{4}. \tag{2.11}$$

Now combining (2.8), (2.9), (2.10), (2.11) yields (2.5), hence the proof is complete. □

2.8 Playing against a fixed strategy

In this section we examine what happens when a player fixes a strategy in a stochastic game. Each player aims to maximize his own individual reward, so it is of interest to see how a player has to play against a fixed strategy.

Definition 2.8.1 *Let $\sigma \in \Sigma$ be a fixed strategy for player 2. Then a strategy $\pi \in \Pi$ for player 1 is called an ε-best reply against σ for initial state $s \in S$, where $\varepsilon \geq 0$, if*

$$\gamma^1_s(\bar{\pi},\sigma) \leq \gamma^1_s(\pi,\sigma)+\varepsilon \qquad \forall \bar{\pi}\in\Pi.$$

The strategy π is called an ε-best reply against σ, if it is an ε-best reply against σ for all initial states $s \in S$. 0-best replies are simply called best replies. Similar definitions hold for the best replies of player 2, and for the discounted reward as well.

First we discuss existence of (ε-)best replies with regard to the average reward.

Theorem 2.8.2

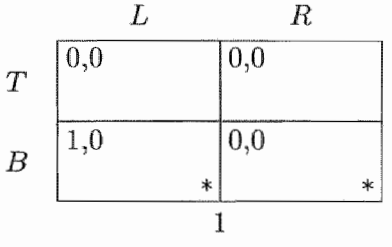
(a) *Against a fixed strategy $\sigma \in \Sigma$, for any $\varepsilon > 0$, player 1 has a pure ε-best reply $\pi \in \Pi^p$. A similar statement holds for player 2 as well.*

(b) *Against any fixed stationary strategy $y \in Y$, player 1 has a pure stationary best reply $i \in I$. A similar statement holds for player 2 as well.*

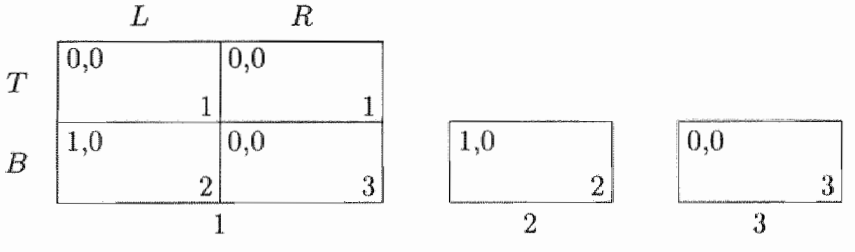
The proof of (a) can be found in Monash [1980] (theorem 1, page 6). Hordijk et al. [1983] showed that if a player has to play against a fixed stationary strategy, then he cannot do better than to play optimally in a related Markov decision process, hence (b) follows from Blackwell [1962].

It is still an open problem whether, against a fixed Markov strategy, pure Markov ε-best replies exist for all $\varepsilon > 0$.
The next example demonstrates that a player does not necessarily have best replies against a fixed strategy.

Example 2.8.3



In order to explain the notation once more, we provide the game in its full form as well.



In state 1, the actions of player 1 are denoted by T (top) and B (bottom), and the actions of player 2 by L (left) and R (right). Consider the Markov strategy g for player 2 which prescribes action L with probability $1 - 1/n$ and action R with probability $1/n$ at stage n. It is clear that, against the strategy g, player 1 is unable to get reward 1 as action L will never be chosen with probability 1. However, for any $\varepsilon > 0$, player 1 can get at least $1 - \varepsilon$ by playing the Markov strategy which prescribes to play action T until $1 - 1/n \geq 1 - \varepsilon$ for the current stage n and to play action B afterwards. This implies that player 1 cannot have best replies against the Markov strategy g. $\triangleleft$

By Hordijk et al. [1983], the following lemma on discounted best replies follows from Blackwell [1962].

Theorem 2.8.4 *Let $\beta \in (0,1)$ and fix a stationary strategy $y \in Y$. Let $\mathcal{B}^1(y)$ denote the set of stationary strategies for player 1 which are β-discounted best replies against y. Then the set $\mathcal{B}^1(y)$ is a nonempty polytope. Moreover, the extreme points of $\mathcal{B}^1(y)$ belong to I, therefore player 1 must have a pure stationary β-discounted best reply $i \in I$ against y. A similar statement holds for player 2 as well.*

2.9 Zero-sum stochastic games and optimality

The theory of zero-sum stochastic games was started by the seminal work of Shapley [1953]. Zero-sum stochastic games are special stochastic games in which the two players have completely opposite interests. These opposite interests are expressed by two assumptions. The first assumption is that

$$r_s^1(i_s, j_s) = -r_s^2(i_s, j_s) \qquad \forall i_s \in I_s,\ \forall j_s \in J_s,\ \forall s \in S,$$

so it may be assumed that the payoffs are payed to player 1 by player 2. Obviously, it also means that the sum of the payoffs of the players is always equal to zero. The second assumption is that

$$\gamma_s^1(\pi, \sigma) = -\gamma_s^2(\pi, \sigma) \qquad \forall \pi \in \Pi,\ \forall \sigma \in \Sigma,\ \forall s \in S,$$

so the sum of the rewards of the players is also equal to zero. When the β-discounted reward is used then the second assumption is of course that

$$\gamma_{\beta s}^1(\pi, \sigma) = -\gamma_{\beta s}^2(\pi, \sigma) \qquad \forall \pi \in \Pi,\ \forall \sigma \in \Sigma,\ \forall s \in S.$$

Note that the second assumption follows from the first one for the discounted reward, but not in the case of the average reward, because player 2 has to use the limit superior instead of the limit inferior in the definition of the average reward (cf. definition 2.5.1) in order to make the sum of the rewards zero. (Recall that all the previously discussed issues remain valid with respect to the limit superior.)
Technically, instead of considering two payoff functions and two rewards, it is easier to consider only player 1's payoffs and player 1's reward and to assume that player 1 tries to maximize his own reward while player 2 tries to minimize player 1's reward. So instead of r^1, γ^1, and γ_β^1 we simply use r, γ, and γ_β.

As the players have completely opposite interests, it is natural to evaluate a strategy for a player by the reward that it guarantees against any strategy of the opponent.

Definition 2.9.1 *In a zero-sum stochastic game, for strategies $\pi \in \Pi$ and $\sigma \in \Sigma$ let*

$$\underline{v}_s(\pi) := \inf_{\sigma' \in \Sigma} \gamma_s(\pi, \sigma') \quad \forall s \in S, \qquad \underline{v}(\pi) := (\underline{v}_s(\pi))_{s \in S}$$

$$\overline{v}_s(\sigma) := \sup_{\pi' \in \Pi} \gamma_s(\pi', \sigma) \quad \forall s \in S, \qquad \overline{v}(\sigma) := (\overline{v}_s(\sigma))_{s \in S}.$$

A strategy π is said to guarantee reward $c_s \in \mathbb{R}$ for initial state $s \in S$, if $\underline{v}_s(\pi) \geq c_s$, and to guarantee $c \in \mathbb{R}^{|S|}$, if $\underline{v}_s(\pi) \geq c_s$ for all $s \in S$. Similarly, a strategy σ is said to guarantee reward $c_s \in \mathbb{R}$ for initial state $s \in S$, if $\overline{v}_s(\sigma) \leq c_s$, and to guarantee $c \in \mathbb{R}^{|S|}$, if $\overline{v}_s(\sigma) \leq c_s$ for all $s \in S$.

Notice that for all $s \in S$, $\pi \in \Pi$, $\sigma \in \Sigma$ we have $\underline{v}_s(\pi) \leq \gamma_s(\pi, \sigma) \leq \overline{v}_s(\sigma)$. This implies

$$\sup_{\pi \in \Pi} \underline{v}_s(\pi) \leq \inf_{\sigma \in \Sigma} \overline{v}_s(\sigma).$$

We now define the solution concepts of zero-sum stochastic games.

Definition 2.9.2 *If there exists a real valued vector $v = (v_s)_{s \in S}$ such that*

$$v_s = \sup_{\pi \in \Pi} \underline{v}_s(\pi) = \inf_{\sigma \in \Sigma} \overline{v}_s(\sigma) \qquad \forall s \in S,$$

then v is called the value of the zero-sum stochastic game.
Assume that the value v exists. Then, for initial state $s \in S$, a strategy $\pi \in \Pi$ is called ε-optimal for player 1, where $\varepsilon \geq 0$, if

$$\underline{v}_s(\pi) \geq v_s - \varepsilon.$$

The strategy π is called ε-optimal, if it is ε-optimal for all initial states $s \in S$. 0-optimal strategies are simply called optimal. Similar definitions hold for player 2 as well.
For $\beta \in (0, 1)$, the β-discounted value v_β and β-discounted optimality are analogously defined.

Note furthermore that, by definition 2.9.2, if the value exists then both players must have ε-optimal strategies for all $\varepsilon > 0$.

Shapley [1953] showed the following result for discounted zero-sum stochastic games.

Theorem 2.9.3 *In any zero-sum stochastic game, for any $\beta \in (0,1)$, the β-discounted value v_β exists and both players have stationary β-discounted optimal strategies.*
Moreover, a stationary strategy $x \in X$ is β-discounted optimal if and only if

$$v_\beta \leq (1-\beta) \cdot r(x,y) + \beta \cdot P(x,y) \cdot v_\beta \qquad \forall y \in Y.$$

A similar statement holds for player 2 as well.

It is fairly appealing in discounted games that optimal strategies of the players can be found in terms of stationary strategies.
Bewley & Kohlberg [1976,I] showed that the discounted values have a unique limit point as the discount factor tends to 1.

Theorem 2.9.4 *In any zero-sum stochastic game, $\lim_{\beta \uparrow 1} v_\beta$ exists.*

Based on deep results of Bewley & Kohlberg [1976,I,II] on discounted values and stationary discounted optimal strategies, Mertens & Neyman [1981] achieved the following fundamental result.

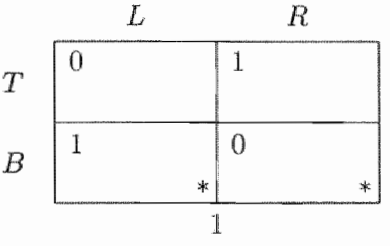
Theorem 2.9.5 *In any zero-sum stochastic game, the value v exists and*

$$v = \lim_{\beta \uparrow 1} v_\beta.$$

Moreover, both players have ε-optimal strategies for all $\varepsilon > 0$.

The fact that the average value is the limit of the β-discounted values, as β tends to 1, is often used in the analysis of stochastic games.
Next we provide an illustration for theorem 2.9.5 by examining the famous stochastic game, called the Big Match, which was introduced by Gillette [1957]. For a long time it was unclear whether the game had a value and whether player 1 had ε-optimal strategies for all $\varepsilon > 0$. The game was only solved 11 years later by Blackwell & Ferguson [1968].

Example 2.9.6 ***The Big Match***



The beauty of the Big Match is that the structure of the game is so simple. In state 1 each player has two actions. Player 1's actions are T and B standing for top and bottom, while player 2's actions are L and R standing for left and right. Action T keeps the play in state 1 with probability 1, while action B leads to an absorbing state, so it ends the game in a strategic sense. Player 1's trouble is that if he uses action B then the place of absorption fully depends on the action chosen by player 2.
We discuss several important issues regarding the Big Match.

Lemma 2.9.7 *The Big Match has the following properties.*

(a) The value for state 1 equals $v_1 = 1/2$.

(b) Player 2 has a stationary optimal strategy $y = (1/2, 1/2) \in Y$.

(c) For $N \in \mathbb{N}$, let π^N be the strategy for player 1 which, for present state 1 and any past history $h \in H$, prescribes action T with probability $1 - (k(h) + N)^{-2}$ and action B with probability $(k(h) + N)^{-2}$, where $k(h)$ denotes the number of stages where player 2 has chosen action R minus the number of stages where player 2 has chosen action L with respect to the history h.

Then for any $\varepsilon > 0$, if $N \in \mathbb{N}$ is sufficiently large, the strategy π^N is an ε-optimal strategy for player 1.

(d) Player 1 has no optimal strategy for initial state 1.

(e) Player 1 has neither stationary nor Markov ε-optimal strategy for initial state 1, if $\varepsilon > 0$ is small. In fact, player 1 can only guarantee reward 0 by stationary strategies and by Markov strategies, namely

$$\sup_{x \in X} \underline{v}_1(x) = \sup_{f \in F} \underline{v}_1(f) = 0.$$

Proof. The proofs of (a),(b), and (c) can be done by showing that, for initial state 1, player 2 can guarantee $1/2$ by playing $y = (1/2, 1/2)$, and, for any $\varepsilon > 0$, player 1 can guarantee $1/2 - \varepsilon$ by the strategies in (c).

It is easy to verify that, for initial state 1, the strategy $y = (1/2, 1/2)$ guarantees $1/2$ for player 2. In fact, regardless of the strategy that player 1 uses against y, the reward always equals $1/2$, since the expected payoff equals $1/2$ for each stage.

For any $\varepsilon > 0$, the strategies in (c) have been found and have been shown to guarantee $1/2 - \varepsilon$ for initial state 1, by Blackwell & Ferguson [1968]. Notice that this strategy is rather complex and player 1 has to make use of the whole past history of the play when choosing his actions.

Now we will prove (d) by showing that no strategy of player 1 can guarantee reward $1/2$ for initial state 1. Take an arbitrary strategy π. Consider the strategy for player 2 which prescribes to play action L as long as π chooses action T with probability 1, to play action R at the first stage when π puts a positive probability on action B, and to play the optimal strategy $(1/2, 1/2)$ afterwards. Then if π always chooses T with probability 1 then the reward is 0. On the other hand, if π ever puts a positive probability on B, say at stage n for the first time, then with a positive probability absorption occurs with payoff zero at stage n, while with the rest of the probability all further expected payoffs equal $1/2$, as player 2 uses $(1/2, 1/2)$. This means that the reward is strictly less than $1/2$ in both cases, so π cannot be optimal for initial state 1. Hence we have shown (d).

Finally, we prove (e). As all the stationary strategies are also Markov strategies, it suffices to show that, for any Markov strategy f for player 1 and for any $\varepsilon > 0$, there exists a strategy σ for player 2 such that $\gamma_1(f, \sigma) \leq \varepsilon$. So take an arbitrary Markov strategy f and an arbitrary $\varepsilon > 0$. We will construct a Markov strategy g for player 2 such that $\gamma_1(f, g) \leq \varepsilon$. Consider the stationary strategy $y = (1, 0)$, which prescribes action L with probability 1. Let $\rho^n(f)$ denote the overall probability that absorption occurs at any of the stages $n+1, n+2, n+3, \ldots$ with respect to f when the initial state is state 1 (clearly, this probability is independent of the strategy used by player 2, due to the transition structure of the game and the fact that f is a Markov strategy). Since the probability that absorption occurs up to stage n converges to $\rho^0(f)$ as n tends to infinity, we have $\rho^0(f) = \lim_{n\to\infty}(\rho^0(f) - \rho^n(f))$, hence $\lim_{n\to\infty} \rho^n(f) = 0$. Then there exists a stage N such that $\rho^N(f) \leq \varepsilon$. Now consider the Markov strategy g for player 2 which prescribes action R up to stage N and action L afterwards. Then, with probability at least $1 - \varepsilon$, either

absorption occurs with payoff zero during the first N stages or entry (T, L) is played at each stage after stage N. Therefore $\gamma_1(f, g) \leq \varepsilon$, which completes the proof of (e). □

2.10 General-sum stochastic games and equilibria

The study of general-sum stochastic games has been started by Fink [1964] and Takahashi [1964]. In general-sum games, in contrast with the previously discussed zero-sum games, the players do not necessarily have strictly opposite interests. The solution concepts value and (ε-)optimal strategies therefore lose their meanings in the context of general-sum games. The most widely applied solution concept here is the concept of (Nash) equilibria. The idea is to find pairs of strategies with the property that neither player could improve his reward individually by choosing another strategy. Such strategy pairs therefore reflect strategically stable situations in the stochastic game.

Definition 2.10.1

(a) *A pair of strategies $(\pi, \sigma) \in \Pi \times \Sigma$ is called a (Nash) ε-equilibrium for initial state $s \in S$, where $\varepsilon \geq 0$, if*

$$\gamma_s^1(\bar{\pi}, \sigma) \leq \gamma_s^1(\pi, \sigma) + \varepsilon \qquad \forall \bar{\pi} \in \Pi$$

$$\gamma_s^2(\pi, \bar{\sigma}) \leq \gamma_s^2(\pi, \sigma) + \varepsilon \qquad \forall \bar{\sigma} \in \Sigma.$$

The strategy pair (π, σ) is an ε-equilibrium, if it is an ε-equilibrium for all initial states $s \in S$. 0-equilibria are simply called equilibria.

(b) *Let $\beta_1, \beta_2 \in (0, 1)$. A pair of strategies $(\pi, \sigma) \in \Pi \times \Sigma$ is called a (Nash) (β_1, β_2)-discounted equilibrium for initial state $s \in S$, if*

$$\gamma_{\beta_1 s}^1(\bar{\pi}, \sigma) \leq \gamma_{\beta_1 s}^1(\pi, \sigma) \qquad \forall \bar{\pi} \in \Pi$$

$$\gamma_{\beta_2 s}^2(\pi, \bar{\sigma}) \leq \gamma_{\beta_2 s}^2(\pi, \sigma) \qquad \forall \bar{\sigma} \in \Sigma.$$

The strategy pair (π, σ) is a (β_1, β_2)-discounted equilibrium, if it is a (β_1, β_2)-discounted equilibrium for all initial states $s \in S$.

When we speak of stationary equilibria or Markov equilibria, we obviously mean equilibria which consist of stationary or Markov strategies, respectively. Note that ε-equilibria can be simply seen as pairs of ε-best replies against each other, with respect to the rewards used by the players.
The following result for the discounted rewards has been proven in Fink [1964] and in Takahashi [1964].

Theorem 2.10.2 *In any stochastic game, there exists a stationary (β_1, β_2)-discounted equilibrium for any $\beta_1, \beta_2 \in (0,1)$.*

The existence of (ε-)equilibria is a much tougher problem for the average reward. The next example demonstrates that 0-equilibria do not necessarily exist, and that stationary and Markov strategies are not always sufficient for achieving ε-equilibria, with small $\varepsilon > 0$.

Example 2.10.3

	L	R
T	0,0	1,−1
B	1,−1 *	0,0 *

1

This game is the general-sum version of the Big Match in example 2.9.6. The only difference is that, here, both players use the limit inferior in the average reward, unlike in example 2.9.6. Assume that the initial state is state 1. As the sum of the payoffs in the game is always zero, we have for all $\pi \in \Pi$ and $\sigma \in \Sigma$ that

$$\gamma_1^1(\pi, \sigma) + \gamma_1^2(\pi, \sigma) \leq 0. \tag{2.12}$$

Suppose that (π, σ) is an ε-equilibrium for initial state 1, where $\varepsilon \geq 0$. In view of lemma 2.9.7-(c), player 1 is able to guarantee that his reward is arbitrarily close to 1/2 regardless of the strategy used by player 2, hence

$$\gamma_1^1(\pi, \sigma) \geq \frac{1}{2} - \varepsilon \tag{2.13}$$

must hold.

Consider the stationary strategy $y = (1/2, 1/2)$ for player 2. Notice that y guarantees reward $-1/2$ to player 2, since his expected payoff is $-1/2$ for any stage, irrespective of the strategy of player 1. Therefore we must have

$$\gamma_1^2(\pi, \sigma) \geq -\frac{1}{2} - \varepsilon. \tag{2.14}$$

By combining (2.12),(2.13),(2.14), we obtain that if (π, σ) is an ε-equilibrium for initial state 1, then

$$\frac{1}{2} - \varepsilon \leq \gamma_1^1(\pi, \sigma) \leq \frac{1}{2} + \varepsilon, \qquad -\frac{1}{2} - \varepsilon \leq \gamma_1^2(\pi, \sigma) \leq -\frac{1}{2} + \varepsilon. \tag{2.15}$$

We will now show that there are no 0-equilibria in this game. Suppose by way of contradiction that (π, σ) is a 0-equilibrium. By (2.15) we have

$$\gamma_1^2(\pi, \sigma) = -\frac{1}{2}.$$

Given the strategy π, construct a strategy $\bar{\sigma}$ for player 2 as in the proof of lemma 2.9.7-(d). Then one can similarly show that $\gamma^2(\pi, \bar{\sigma}) > -1/2$, hence (π, σ) cannot be a 0-equilibrium.
Next, we show that there are no ε-equilibria in terms of stationary or Markov strategies for any $\varepsilon \in [0, 1/6)$. As all stationary strategies are Markov strategies as well, it suffices to prove the statement for Markov strategies. Take any $\varepsilon \in [0, 1/6)$ and suppose by way of contradiction that (f, g) is a Markov ε-equilibrium. The inequalities in (2.15) imply that

$$\gamma_1^2(f, g) \leq -\frac{1}{2} + \varepsilon < -\frac{1}{3}. \tag{2.16}$$

Given the strategy f and the chosen ε, construct a Markov strategy $\bar{g}$ for player 2 as in the proof of lemma 2.9.7-(e). Then, by using analogous arguments, one can verify that $\gamma_1^2(f, \bar{g}) \geq -\varepsilon$. Now the choice of ε yields $\gamma_1^2(f, \bar{g}) > -1/6$. Thus (2.16) implies

$$\gamma_1^2(f, \bar{g}) > \gamma_1^2(f, g) + \varepsilon,$$

hence (f, g) cannot be an ε-equilibrium.
Finally, we wish to mention that, for all $\varepsilon > 0$, the strategies in lemma 2.9.7-(b),(c) form an ε-equilibrium in this game. Note, however, that player 1 needs to use a history dependent strategy in these ε-equilibria. $\triangleleft$

The existence of ε-equilibria in terms of history dependent strategies has been an open problem in stochastic game theory for a long time. Although this

question had been previously answered in the affirmative for several classes of stochastic games, it was only recently that Vieille [1997,I,II] derived the existence of ε-equilibria for all two-person stochastic games, for all $\varepsilon > 0$. The existence problem, however, is still open in stochastic games with more than two players.

For zero-sum stochastic games, in view of theorem 2.9.5, we have $v = \lim_{\beta \uparrow 1} v^{\beta}$. For general-sum stochastic games the relation between the discounted and the average game is much weaker, as it may happen that the rewards corresponding to stationary discounted equilibria are far away from the rewards that correspond to average ε-equilibria, for small $\varepsilon > 0$. This is fully clarified by the game in Sorin [1986].

Example 2.10.4

	L	R
T	0,1	1,0
B	1,0 *	0,2 *

1

Assume that the initial state is state 1. Sorin [1986] showed that the set of rewards which, for all $\varepsilon > 0$, correspond to ε-equilibria is

$$L = \text{conv}\left\{(1/2, 1), (2/3, 2/3)\right\},$$

where conv stands for the convex hull of a set. On the other hand, the set of rewards corresponding to (β_1, β_2)-discounted equilibria is the same singleton for all $\beta_1, \beta_2 \in (0, 1)$:

$$L_{\text{disc}} = \left\{(1/2, 2/3)\right\}.$$

Therefore there is a gap between the average solutions and the discounted solutions in this game. $\triangleleft$

We would like to stress that, despite the above example, the discounted solutions are frequently used in the analysis of equilibria for the average reward. Finally, we wish to mention that the concept of (ε-)equilibria naturally extends to stochastic games where the players use different rewards. In this sense we can speak of (ε-)equilibria in zero-sum stochastic games as well (recall that,

in the case of the average reward, player 2 uses the limit superior instead of the limit inferior). It is easy to see that, in zero-sum stochastic games, for any $\varepsilon > 0$, any pair of ε-optimal strategies forms a 2ε-equilibrium, and any ε-equilibrium must be formed by 2ε-optimal strategies.

2.11 Special classes of stochastic games

In this section we give a brief overview of the most important classes of stochastic games with the most important issues.

Definition 2.11.1 *A stochastic game is called*

(a) *a unichain stochastic game, if there is only one ergodic set of states with respect to any pair of stationary strategies.*

(b) *a perfect information stochastic game, if S can be partitioned into S^1 and S^2 such that $|J_s| = 1$ for all $s \in S^1$ and $|I_s| = 1$ for all $s \in S^2$.*

(c) *a switching control stochastic game, if S can be partitioned into S^1 and S^2 such that $p_s(i_s, j_s)$ is independent of j_s for all $i_s \in I_s$, $s \in S^1$, and $p_s(i_s, j_s)$ is independent of i_s for all $j_s \in J_s$, $s \in S^2$.*

(d) *a stochastic game with additive reward and additive transition (ARAT) structure, if $r_s^k(i_s, j_s)$ and $p_s(i_s, j_s)$ can be decomposed as*

$$r_s^k(i_s, j_s) = r_s^{k,1}(i_s) + r_s^{k,2}(j_s), \qquad p_s(i_s, j_s) = p_s^1(i_s) + p_s^2(j_s)$$

for all $i_s \in I_s$, $j_s \in J_s$, $s \in S$.

(e) *a stochastic game with state independent transitions (SIT), if the cardinality of the action spaces is independent of the state and, assuming that $I_s = I_t =: \bar{I}$ and $J_s = J_t =: \bar{J}$ for all $s, t \in S$, it holds that $p_s = p_t$ for all $s, t \in S$.*

(f) *a stochastic game with separable rewards and state independent transitions (SER-SIT), if it is a SIT stochastic game and, assuming that $I_s = I_t =: \bar{I}$ and $J_s = J_t =: \bar{J}$ for all $s, t \in S$, the function $r_s^k(i, j)$ can be decomposed as*

$$r_s^k(i, j) = c_s^k + d^k(i, j)$$

for all $s \in S$, $i \in \bar{I}$, $j \in \bar{J}$.

(g) *a repeated game with absorbing states, if all the states but one are absorbing.*

(h) *a recursive stochastic game, if the payoffs are equal to zero in all non-absorbing states.*

For the zero-sum case the following theorem summarizes the most important results.

Theorem 2.11.2

(a) *In zero-sum unichain stochastic games, both players have stationary optimal strategies, and the value is independent of the initial state (cf. Hoffman & Karp [1966], Thuijsman [1992]).*

(b) *In zero-sum perfect information stochastic games, both players have pure stationary optimal strategies (cf. Liggett & Lippman [1969]).*

(c) *In zero-sum switching control stochastic games, both players have stationary optimal strategies (cf. Filar [1981]).*

(d) *In zero-sum ARAT stochastic games, both players have pure stationary optimal strategies (cf. Raghavan et al. [1985]).*

(e) *In zero-sum SIT stochastic games, both players have stationary optimal strategies, and the value is independent of the initial state (cf. Thuijsman [1992]).*

(f) *In zero-sum SER-SIT stochastic games, both players have stationary optimal strategies such that the prescribed mixed actions are state independent, and the value is independent of the initial state (cf. Parthasarathy et al. [1984]).*

(g) *In zero-sum repeated games with absorbing states, both players have ε-optimal strategies for all $\varepsilon > 0$ (cf. Kohlberg [1974]).*

(h) *In zero-sum recursive stochastic games, both players have stationary ε-optimal strategies for all $\varepsilon > 0$ (cf. Everett [1957], Thuijsman & Vrieze [1992]).*

In the general-sum case the following results are known.

Theorem 2.11.3

(a) *In unichain stochastic games, stationary equilibria exist (cf. Rogers [1969], Sobel [1971], Federgruen [1978], Thuijsman [1992]).*

(b) *In perfect information stochastic games, pure equilibria exist (cf. Liggett & Lippman [1969], Thuijsman & Raghavan [1997]).*

(c) *In switching control stochastic games, ε-equilibria exist for all $\varepsilon > 0$ (cf. Thuijsman & Raghavan [1997]).*

(d) *In ARAT stochastic games, pure equilibria exist (cf. Thuijsman & Raghavan [1997].)*

(e) *In SIT stochastic games, ε-equilibria exist for all $\varepsilon > 0$ (cf. Thuijsman [1992]).*

(f) *In SER-SIT stochastic games, stationary equilibria exist with the property that the prescribed mixed actions are independent of the state (cf. Parthasarathy et al. [1984]).*

(g) *In repeated games with absorbing states, ε-equilibria exist for all $\varepsilon > 0$ (cf. Vrieze & Thuijsman [1989]).*

Recently, Vieille derived that, for all $\varepsilon > 0$, ε-equilibria exist in all two-person stochastic games. In order to achieve this result, first he proved that the general existence problem in two-person stochastic games reduces to the existence problem in a specific class of two-person recursive games (cf. Vieille [1994] and [1997,I]), and afterwards he showed that this specific class of two-person recursive games possesses ε-equilibria for all $\varepsilon > 0$ (cf. Vieille [1997,II]).

Part I

Zero-sum stochastic games

Chapter 3

Simplifying optimal strategies

3.1 Introduction

In zero-sum stochastic games, by theorem 2.9.5, the value exists and the players have ε-optimal strategies for all $\varepsilon > 0$. However, optimal strategies need not always exist, as illustrated by the Big Match in example 2.9.6 and lemma 2.9.7-(d). In this chapter, which is mainly based on Flesch et al. [1998,I], we examine how optimal strategies can be simplified when they exist.
First, we state the main result of this chapter, which will follow from theorem 3.3.1 below.

Main Theorem 3 *In a zero-sum stochastic game, if a player has an optimal strategy then he has stationary ε-optimal strategies, for all $\varepsilon > 0$, and he has Markov optimal strategies as well.*

In other words, we show that optimal strategies, when they exist, can be simplified by stationary ε-optimal strategies, for all $\varepsilon > 0$, and by Markov optimal strategies as well. So instead of playing a complex history dependent optimal strategy, the player can also play a stationary ε-optimal strategy with an arbitrary small $\varepsilon > 0$, or he can even achieve optimality in the class of Markov strategies. We present such a construction for which we do not even need to know any optimal strategy, which makes the result even stronger.
We provide two examples showing the sharpness of the result. Example 3.3.2 will demonstrate that the existence of stationary optimal strategies is not implied by the existence of optimal strategies, while example 3.4.6 will clarify that the existence of stationary ε-optimal strategies, for all $\varepsilon > 0$, is not

sufficient for the existence of optimal strategies.
In several stochastic games, it is easy to show that stationary ε-optimal strategies do not exist for small $\varepsilon > 0$. In such games, by the above theorem, we may exclude the existence of optimal strategies as well. Note that, without knowing this result, it would be a much harder problem to check the existence of optimal strategies, since optimal strategies could only exist in terms of history dependent strategies.
The above theorem moreover provides a sufficient condition for the existence of stationary ε-optimal strategies. For many classes of stochastic games, where on the payoff and transition structures special conditions are imposed, stationary ε-optimal strategies exist for all $\varepsilon > 0$ (cf. theorem 2.11.2). Here, instead of providing such structural conditions, the existence of optimal strategies will be proven to be sufficient.
At the end of this chapter we make several remarks regarding the construction and the proofs. There we also treat possible simplifications of strategies that are only optimal for particular initial states.

3.2 Preliminaries

The following lemma was shown by von Neumann [1928].

Lemma 3.2.1 *Let $s \in S$ and let $c_s : X_s \times Y_s \longmapsto \mathbb{R}$ be linear in both components. Then there exist $x_s \in X_s$, $y_s \in Y_s$, and a unique $C_s \in \mathbb{R}$ such that*

$$c_s(x_s, y_s') \geq C_s \geq c_s(x_s', y_s) \qquad \forall x_s' \in X_s,\ \forall y_s' \in Y_s.$$

Lemma 3.2.2 *Let $s \in S$. Let c_s and C_s be as in lemma 3.2.1. Then the sets*

$$O_s^1 := \{x_s \in X_s |\, c_s(x_s, y_s') \geq C_s \quad \forall y_s' \in Y_s\}$$

$$O_s^2 := \{y_s \in Y_s |\, c_s(x_s', y_s) \leq C_s \quad \forall x_s' \in X_s\}$$

$$\bar{O}_s^1 := \{x_s \in X_s |\, c_s(x_s, y_s') = C_s \quad \forall y_s' \in O_s^2\}$$

$$\bar{O}_s^2 := \{y_s \in Y_s |\, c_s(x_s', y_s) = C_s \quad \forall x_s' \in O_s^1\}$$

are nonempty polytopes. Furthermore, if $\bar{I}_s$ and $\bar{J}_s$ denote the extreme points of the sets $\bar{O}_s^1$ and $\bar{O}_s^2$, respectively, then

$$\bar{I}_s = \{i_s \in I_s |\, \exists x_s \in O_s^1 :\ x_s(i_s) > 0\}$$

$$\bar{J}_s = \{j_s \in J_s |\, \exists y_s \in O_s^2 :\ y_s(j_s) > 0\}.$$

Proof. The above sets O_s^1, O_s^2, $\bar{O}_s^1$, $\bar{O}_s^2$ are nonempty by lemma 3.2.1. One can also show that these sets are polytopes by using the linearity of c_s in both components, which can be found in most elementary books on the theory of matrix games. The last part of the statement is shown in Gale & Sherman [1950] and in Bohnenblust et al. [1950]. □

Definition 3.2.3 *For $s \in S$, $x_s \in X_s$, $y_s \in Y_s$ let*

$$V_s(x_s, y_s) := \sum_{t \in S} p_s(t|x_s, y_s) \cdot v_t.$$

For $x \in X$, $y \in Y$ let

$$V(x, y) := (V_s(x_s, y_s))_{s \in S}.$$

Here $V_s(x_s, y_s)$ is the expectation of the value after transition from state s with regard to the pair of mixed actions (x_s, y_s).
The following lemma intuitively says that player 1 can guarantee that the value does not decrease in expectation after transition, while player 2 can make sure that the value does not increase in expectation after transition.

Lemma 3.2.4 *For any $s \in S$, there exist $x_s \in X_s$ and $y_s \in Y_s$ such that*

$$V_s(x_s, y_s') \geq v_s \geq V_s(x_s', y_s) \qquad \forall x_s' \in X_s,\ \forall y_s' \in Y_s.$$

Proof. Let $s \in S$. In view of lemma 3.2.1, there exist $x_s \in X_s$, $y_s \in Y_s$, and a unique $C_s \in \mathbb{R}$ such that

$$V_s(x_s, y_s') \geq C_s \geq V_s(x_s', y_s) \qquad \forall x_s' \in X_s,\ \forall y_s' \in Y_s.$$

So we have to show that $v_s = C_s$. Assume by way of contradiction that $v_s > C_s$; the proof is similar when $v_s < C_s$ is assumed.
Let

$$d := v_s - C_s > 0.$$

We derive a contradiction by showing that player 1 does not have ε-optimal strategies for initial state s, for any $\varepsilon \in (0, d)$. Let $\varepsilon \in (0, d)$ and take an arbitrary strategy $\pi \in \Pi$. Recall that the mixed action prescribed by π for stage 1 in the initial state s is denoted by π_s. Consider a strategy σ^ε for player 2 which prescribes to play y_s at stage 1 in state s and to play a $((d - \varepsilon)/2)$-optimal strategy afterwards. Then

$$V_s(\pi_s, y_s) \leq C_s = v_s - d.$$

From stage 2 on player 2 plays a $((d-\varepsilon)/2)$-optimal strategy, so

$$\begin{aligned}\gamma_s(\pi,\sigma^\varepsilon) &\leq \sum_{t\in S} p_s(t|\pi_s,y_s)\cdot\left(v_t+\frac{d-\varepsilon}{2}\right)\\ &= V_s(\pi_s,y_s)+\frac{d-\varepsilon}{2}\\ &\leq (v_s-d)+\frac{d-\varepsilon}{2}\\ &< v_s-\varepsilon.\end{aligned}$$

Thus player 1 cannot have an ε-optimal strategy for initial state s, which is a contradiction. □

We will deal with restricted games derived from the original game Γ. Assume that $S' \subset S$ is a non-empty set of states and $X'_s \subset X_s$, $Y'_s \subset Y_s$ are nonempty polytopes for all $s \in S'$. Suppose that all pairs of mixed actions in $X'_s \times Y'_s$, for any $s \in S'$, only induce transitions to states in S'. Then we may define a restricted game Γ', derived from the original game Γ, where the state space is S' and the players are restricted to using mixed actions in X'_s and Y'_s, if the play is in any state $s \in S'$. In the restricted game Γ', let H' denote the set of finite histories, Π' and Σ' the sets of (history dependent) strategies, and γ' the average reward. The stationary strategy spaces in Γ' are $X' := \times_{s\in S'} X'_s$ and $Y' := \times_{s\in S'} Y'_s$.
For the restricted game Γ', for any $\beta \in (0,1)$, it can be shown similarly to theorem 2.9.3 that the β-discounted value v'_β exists and both players have stationary β-discounted optimal strategies. Moreover, $x \in X'$ is β-discounted optimal in Γ' if and only if

$$v'_\beta \leq (1-\beta)\cdot r(x,y)+\beta\cdot P(x,y)\cdot v'_\beta \qquad \forall y \in Y'. \tag{3.1}$$

Theorem 2.9.4 applies for Γ' as well, so $\lim_{\beta\uparrow 1} v'_\beta$ exist. Let

$$v' := \lim_{\beta\uparrow 1} v'_\beta. \tag{3.2}$$

Note that we do not claim that v' is the average value of Γ' as in theorem 2.9.5 for the original game, because the players only observe pure actions, which do not necessarily correspond to extreme points of the restricted spaces of mixed actions. However one can show, by using an appropriate sequence of discount factors as in Mertens & Neyman [1981], that, against any fixed strategy in Π',

for any $\varepsilon > 0$ player 2 can make sure that player 1's reward is at most $v' + \varepsilon$, namely

$$\sup_{\pi \in \Pi'} \inf_{\sigma \in \Sigma'} \gamma'_s(\pi, \sigma) \leq v'_s \qquad \forall s \in S'. \tag{3.3}$$

3.3 The construction

Let

$$X^*_s := \{x_s \in X_s | \, V_s(x_s, y_s) \geq v_s \ \ \forall y_s \in Y_s\} \quad \forall s \in S, \qquad X^* := \times_{s \in S} X^*_s,$$

$$Y^*_s := \{y_s \in Y_s | \, V_s(x_s, y_s) = v_s \ \ \forall x_s \in X^*_s\} \quad \forall s \in S, \qquad Y^* := \times_{s \in S} Y^*_s.$$

Note the asimilarity in the definitions of X^*_s and Y^*_s, $s \in S$. In view of lemmas 3.2.4 and 3.2.2, for all $s \in S$, the sets X^*_s, Y^*_s are nonempty polytopes and there exists a $J^*_s \subset J_s$ such that $Y^*_s = \text{conv}(J^*_s)$, where conv stands for the convex hull of a set. Let

$$J^* := \times_{s \in S} J^*_s.$$

As in section 3.2, we may define a restricted game Γ^*, derived from the original game Γ, where the state space is S and the players are restricted to using mixed actions in X^*_s and Y^*_s, if the play is in any state $s \in S$. In the restricted game Γ^*, the sets of stationary strategies are X^* and Y^*.

By the finiteness of the state and action spaces, there exists a countable subset of discount factors $\mathcal{B} \subset (0, 1)$ such that 1 is a limit point of $\mathcal{B}$ and there are stationary β-discounted optimal strategies $x_\beta \in X^*$ in the restricted game Γ^* such that the sets

$$\{i_s \in I_s | \, x_{\beta s}(i_s) > 0\}, \qquad s \in S,$$

are independent of $\beta \in \mathcal{B}$. In this chapter, each time that we are dealing with discount factors, discounted optimal strategies, or with limits when the discount factors converge to 1, we will have such a subset of discount factors $\mathcal{B}$ in mind.

If Z is a polytope then let Relint(Z) denote the relative interior of the polytope Z, which is defined as the set of points in Z which can be written as a convex combination of all the extreme points of Z with only strictly positive coefficients.

Theorem 3.3.1 *Assume that player 1 has an optimal strategy in the game* Γ.

(a) *For any* $\beta \in \mathcal{B}$, *let* $x_\beta \in X^*$ *be a* β*-discounted optimal strategy in the restricted game* Γ^* *and let* $x \in Relint(X^*)$. *Then, for any* $\varepsilon > 0$, *if* $\beta \in \mathcal{B}$, $\tau \in (0,1)$ *are sufficiently large then the stationary strategy*

$$x_\beta^\tau := \tau \cdot x_\beta + (1-\tau) \cdot x \in X^*$$

is ε*-optimal in the game* Γ.

(b) *Let* $\varepsilon_n > 0$, $n \in \mathbb{N}$, *be a strictly monotonously decreasing sequence which converges to* 0. *Let* $x_n \in X^*$ *be* ε_n*-optimal in* Γ *for all* $n \in \mathbb{N}$. *Then there exist a sequence* K_n *in* $\mathbb{N}$ *such that the Markov strategy* f *which prescribes to play* x_1 *for the first* K_1 *stages, then to play* x_2 *for the next* K_2 *stages, and so on, is optimal in the game* Γ.

A similar statement holds for player 2 as well.

We provide an illustration of the above constructions with the help of the following example. This example also demonstrates that the existence of optimal strategies does not yield the existence of stationary optimal strategies.

Example 3.3.2

	L	R
T	0	2 $*$
B	1	0

1

The value for initial state 1 is $v_1 = 1$. It is not hard to show that there are optimal strategies for player 1 (later we will construct optimal Markov strategies).
Assume that the initial state is state 1. Following the construction for stationary ε-optimal strategies, we have $X^* = X$, $Y^* = \{(1,0)\}$. Now the unique β-discounted optimal strategy of player 1 in Γ^* is $x_\beta = (0,1)$ for all $\beta \in (0,1)$. The role of x_β is to play well as long as player 2 plays in the restricted game Γ^*, namely to guarantee the value v as long as player 2 chooses action L in state 1. However, an enforcement is needed to make sure that player 2 is not

better off by playing outside Y^*, namely by choosing action R. Therefore we take a strategy $x \in \text{Relint}(X^*)$, for example $x = (1/2, 1/2)$, which will force player 2 not to choose action R, since then R leads to absorption with payoff 2. Now for $\tau, \beta \in (0,1)$ we have

$$x_\beta^\tau = \tau \cdot x_\beta + (1-\tau) \cdot x = (1/2 - \tau/2, 1/2 + \tau/2).$$

The strategy x_β^τ is ε-optimal for large τ and β indeed, as the stationary strategies $(p, 1-p)$ are ε-optimal for all $p \in (0, \varepsilon]$.
Note that player 1 has no stationary optimal strategy in this game. One can argue as follows. If a stationary strategy x prescribes action T with a positive probability then x only gives a reward strictly less then 1 if player 2 always chooses action L. On the other hand, if x chooses action B with probability 1, then if player 2 always takes action R then the reward is 0. Thus no stationary strategy can guarantee $v_1 = 1$.
Nevertheless, a Markov optimal strategy can be constructed as in theorem 3.3.1-(b). The idea is to increase β and τ simultaneously during the play so that player 1 plays better and better in the restricted game. However, τ must be increased sufficiently slowly so that player 2 cannot choose action R "too often" without absorption. Formally, let $\varepsilon_n = 1/n$ and take the stationary ε_n-optimal strategy $x_n = (\varepsilon_n, 1-\varepsilon_n) \in X^*$ for all $n \in \mathbb{N}$. Choose, for instance, $K_n = 1$ for all $n \in \mathbb{N}$. Now, let f be the Markov strategy as in theorem 3.3.1-(b). So at stage n, the strategy f chooses action T with probability $1/n$ and action B with probability $1 - 1/n$. One can verify that f is optimal. We only give an intuitive argument. If player 2 chooses action R with a "positive frequency" then absorption occurs with probability 1 due to the slowly decreasing probabilities on action T; while almost always choosing action L yields reward 1 since the probabilities on action B converge to 1. (A rigorous proof for the optimality of the Markov strategy f can be given by using techniques as in chapter 5). $\triangleleft$

3.4 The proof

In the restricted game Γ^*, let H^* denote the set of finite histories, Π^* and Σ^* the sets of (history dependent) strategies, and γ^* the average reward. Note that $H^* \subset H$. Let v_β^* denote the β-discounted value for the restricted game Γ^*, and let

$$v^* := \lim_{\beta \uparrow 1} v_\beta^*.$$

We also introduce

$$\bar{\Pi} := \{\pi \in \Pi | \, \pi_s(h) \in X_s^* \quad \text{for all } s \in S \text{ and } h \in H^*\}$$

$$\bar{\Sigma} := \{\sigma \in \Sigma | \, \sigma_s(h) \in Y_s^* \quad \text{for all } s \in S \text{ and } h \in H^*\};$$

so $\bar{\Pi}$ and $\bar{\Sigma}$ are the sets of strategies in the original game Γ which behave as strategies in Π^* and Σ^* as long as the play is in the restricted game Γ^*. Each stationary strategy in X^* and Y^* can be naturally seen as a stationary strategy in $\bar{\Pi}$ and $\bar{\Sigma}$.

The following lemma clarifies why the sets X^* and Y^* play an important role when player 1 has an optimal strategy in the original game Γ. This lemma states that if π is an optimal strategy for player 1 in Γ then, for any present state $t \in S$ and past history with a positive occurrence probability with respect to (π, σ) for some $\sigma \in \bar{\Sigma}$, the strategy π prescribes a mixed action belonging to X_t^*. In other words, if player 2 uses a strategy $\sigma \in \bar{\Sigma}$ then the optimal strategy π will behave as a strategy in $\bar{\Pi}$.

Lemma 3.4.1 *Let $\pi \in \Pi$ be an optimal strategy for player 1 in the game Γ. Let $s \in S$ and let*

$$U_s := \{(h,t) \in H_s \times S | \, \exists \sigma \in \bar{\Sigma} : \quad \mathcal{P}_{s\pi\sigma}(h) > 0 \quad \textit{and} \quad \mathcal{P}_{s\pi\sigma}(t|h) > 0\},$$

where $\mathcal{P}_{s\pi\sigma}(t|h)$ is the probability that, with respect to (π, σ), state t becomes the new state after history h.
Then $h \in H_s^*$ and $\pi_t(h) \in X_t^*$ for all pairs $(h,t) \in U_s$.

Proof. Suppose the opposite. Then, there is a shortest history $\bar{h}^n \in H_s$, say up to stage n, with the following properties: there exist a $\sigma \in \bar{\Sigma}$ and a state $t \in S$ such that

$$\mathcal{P}_{s\pi\sigma}(\bar{h}^n) > 0 \quad \text{and} \quad \mathcal{P}_{s\pi\sigma}(t|\bar{h}^n) > 0,$$

and

$$\bar{h}^n \in H_s \setminus H_s^* \quad \text{or} \quad \pi_t(\bar{h}^n) \in X_t \setminus X_t^*.$$

First we show that $\bar{h}^n \in H_s^*$ must hold. If $n = 0$ then, clearly, $\bar{h}^0 \in H_s^*$. Assume now that $n \geq 1$. Let $\bar{h}^{n-1}$ denote the history $\bar{h}^n$ up to stage $n-1$ and s^n the state for stage n in $\bar{h}^n$. Then $(\bar{h}^{n-1}, s^n) \in U_s$ with the very same σ. Because $\bar{h}^n$ is a shortest history with the above properties, we must have

$$\bar{h}^{n-1} \in H_s^*, \qquad \pi_{s^n}(\bar{h}^{n-1}) \in X_{s^n}^*.$$

Since $\sigma \in \bar{\Sigma}$, we also obtain

$$\sigma_{s^n}(\bar{h}^{n-1}) \in Y^*_{s^n},$$

therefore $\bar{h}^n \in H^*_s$ must hold indeed.
Because $\bar{h}^n \in H^*_s$, the properties of $\bar{h}^n$ imply $\pi_t(\bar{h}^n) \in X_t \setminus X^*_t$. Therefore, there exists a $\bar{y}_t \in Y_t$ such that

$$\tau := v_t - V_t(\pi_t(\bar{h}^n), \bar{y}_t) > 0.$$

By lemma 3.2.4, there also exists a $y \in Y$ such that $V_z(x_z, y_z) \leq v_z$ for all $x_z \in X_z$ and $z \in S$.
Let $s^1 := s$, and let s^m, $m \geq 2$, denote the random variable for the state at stage m, and let θ^m denote random variable for the history up to stage $m \in \mathbb{N}$. Let

$$\delta \in \left(0, \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) \cdot \tau\right).$$

Let $\sigma^\delta \in \Sigma$ be the strategy that prescribes to play as follows: play σ during the first n stages; at stage $n+1$, if $\theta^n = \bar{h}^n$ and $s^{n+1} = t$ then play $\bar{y}_t$ while if $\theta^n \neq \bar{h}^n$ or $s^{n+1} \neq t$ then play the mixed action $y_{s^{n+1}}$; and finally, play a δ-best reply against $\pi[\theta^{n+1}]$ from stage $n+2$ on. Note that

$$\mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) = \mathcal{P}_{s^1\pi\sigma^\delta}(\bar{h}^n) > 0, \qquad \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) = \mathcal{P}_{s^1\pi\sigma^\delta}(t|\bar{h}^n) > 0.$$

Since $\bar{h}^n$ is a shortest history with its above specified properties, by the definitions of X^* and Y^* we have

$$\mathcal{E}_{s^1\pi\sigma^\delta}(v_{s^{n+1}}) = v_{s^1},$$

and by the used mixed actions at stage $n+1$

$$\mathcal{E}_{s^1\pi\sigma^\delta}(v_{s^{n+2}}) \leq \mathcal{E}_{s^1\pi\sigma^\delta}(v_{s^{n+1}}) - \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) \cdot \tau.$$

From stage $n+2$ player 2 plays a δ-best reply, so the choice of δ yields

$$\begin{aligned}
\gamma_{s^1}(\pi, \sigma^\delta) &\leq \mathcal{E}_{s^1\pi\sigma^\delta}(v_{s^{n+2}}) + \delta \\
&\leq \mathcal{E}_{s^1\pi\sigma^\delta}(v_{s^{n+1}}) - \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) \cdot \tau + \delta \\
&= v_{s^1} - \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) \cdot \tau + \delta \\
&< v_{s^1},
\end{aligned}$$

which contradicts the optimality of π. □

The condition that player 1 has an optimal strategy in Γ is only needed for the next lemma. We show that, if player 1 has an optimal strategy in Γ, then he can guarantee a reward at least v in the restricted game Γ^*. This is based on the facts that optimal strategies of player 1 guarantee a reward at least the value v in the original game Γ and, in view of the previous lemma, they can only prescribe mixed actions in X_s^*, if the play is in any state s, against any strategy of player 2 in $\bar{\Sigma}$. The second part of the statement says that player 1 cannot guarantee more than the limit of the β-discounted values in Γ^*.

Lemma 3.4.2 *Suppose that player 1 has an optimal strategy $\pi \in \Pi$. Then*

$$v_s \leq \sup_{\pi^* \in \Pi^*} \inf_{\sigma^* \in \Sigma^*} \gamma_s^*(\pi^*, \sigma^*) \leq v_s^* \qquad \forall s \in S.$$

Proof. The second inequality follows from (3.3), so we only have to show the first one.
For any $s \in S$, let U_s be as in lemma 3.4.1. Take an arbitrary $x \in X^*$. By using lemma 3.4.1, we may define a strategy $\pi^* \in \Pi^*$ as follows: for $s \in S$, $h \in H_s^*$ and $t \in S$ let

$$\pi_t^*(h) := \begin{cases} \pi_t(h) & \text{if } (h,t) \in U_s \\ x_t & \text{if } (h,t) \notin U_s. \end{cases}$$

Let $\sigma^* \in \Sigma^*$ be arbitrary. Choose an arbitrary $\sigma \in \bar{\Sigma}$ such that

$$\sigma_t(h) = \sigma_t^*(h) \qquad \forall t \in S,\ \forall h \in H^*.$$

Then by the optimality of π and by lemma 3.4.1, we have

$$v_s \leq \gamma_s(\pi, \sigma) = \gamma_s^*(\pi^*, \sigma^*) \qquad \forall s \in S.$$

Since $\sigma^* \in \Sigma^*$ was arbitrary, the first inequality of the lemma also follows. So the proof is complete. □

The next result shows the effectiveness of the β-discounted optimal strategies in the restricted game Γ^*.

Lemma 3.4.3 *Let $\varepsilon > 0$. For $\beta \in \mathcal{B}$, let $x_\beta \in X^*$ be a β-discounted optimal strategy of player 1 in Γ^*, and let $y \in Y^*$. Suppose that $E \subset S$ is a closed set of states with respect to (x_β, y) for all $\beta \in \mathcal{B}$. Then for large $\beta \in \mathcal{B}$*

$$\gamma_s(x_\beta, y) \geq \min_{t \in E} v_t^* - \varepsilon \qquad \forall s \in E.$$

Proof. Using inequality (3.1) for Γ^* we have

$$(1-\beta)\cdot r(x_\beta, y) + \beta \cdot P(x_\beta, y)\cdot v^*_\beta \geq v^*_\beta \qquad \forall \beta \in \mathcal{B}.$$

By (2.1), multiplying this inequality by $Q(x_\beta, y)$ yields

$$Q(x_\beta, y)\cdot r(x_\beta, y) \geq Q(x_\beta, y)\cdot v^*_\beta \qquad \forall \beta \in \mathcal{B}.$$

Let $s \in E$ be arbitrary. Then

$$\sum_{t\in E} q_s(t|x_\beta, y)\, r_t(x_{\beta t}, y_t) \geq \sum_{t\in E} q_s(t|x_\beta, y)\, v^*_{\beta t} \qquad \forall \beta \in \mathcal{B}.$$

Since $s \in E$, the closedness of E implies that if $q_s(t|x_\beta, y) > 0$ then $t \in E$. Hence for large $\beta \in \mathcal{B}$, by using theorem 2.7.1-(a) and the definition of v^*, we must have

$$\begin{aligned}
\gamma_s(x_\beta, y) &= \sum_{t\in E} q_s(t|x_\beta, y)\, r_t(x_{\beta t}, y_t) \\
&\geq \sum_{t\in E} q_s(t|x_\beta, y)\, v^*_{\beta t} \\
&\geq \sum_{t\in E} q_s(t|x_\beta, y)\, (v^*_t - \varepsilon) \\
&\geq \min_{t\in E} v^*_t - \varepsilon,
\end{aligned}$$

so the proof is complete. □

Next, we discuss some properties of stationary strategies belonging to X^*.

Lemma 3.4.4 *Let $x \in X^*$ and $y \in Y$. Suppose E is an ergodic set with respect to (x, y). Then $v_s = v_t$ for all $s, t \in E$. Furthermore, if $x \in Relint(X^*)$ then necessarily $y_s \in Y^*_s$ for all $s \in E$.*

Proof. By $x \in X^*$ and by the closedness of E for (x, y) we obtain

$$v_s \leq V_s(x_s, y_s) = \sum_{t\in E} p_s(t|x_s, y_s)\, v_t \qquad \forall s \in E.$$

Let $\bar{E} := \{s \in E |\, v_s = \max_{t\in E} v_t\}$. The above inequalities imply that $\bar{E}$ is a closed set of states for (x, y), so since E is an ergodic set for (x, y), we have $\bar{E} = E$. Therefore $v_s = v_t =: v_E$ for all $s, t \in E$.

Now suppose that $x \in \text{Relint}(X^*)$. Then $(\bar{x}_s, y_s)$ only induces transitions to states in E for any $\bar{x}_s \in X_s^*$, $s \in E$, hence

$$V_s(\bar{x}_s, y_s) = \sum_{t \in E} p_s(t|\bar{x}_s, y_s)\, v_E = v_E = v_s \qquad \forall \bar{x}_s \in X_s^*,\ \forall s \in E,$$

which implies that $y_s \in Y_s^*$ for all $s \in E$. □

An important property of convex combinations of stationary strategies is stated in the next lemma.

Lemma 3.4.5 *For $\tau \in (0,1)$, $x^1, x^2 \in X$ let $x^\tau := \tau x^1 + (1-\tau)x^2$. Suppose that E is an ergodic set with respect to (x^τ, y) for some $y \in Y$. Let $\varepsilon > 0$ and $d \in \mathbb{R}$. If*

$$\gamma_s(x^1, y) \geq d \qquad \forall s \in E$$

then for sufficiently large $\tau \in (0,1)$

$$\gamma_s(x^\tau, y) \geq d - \varepsilon \qquad \forall s \in E.$$

Proof. Let $\delta \in (0,1)$. As $\gamma_s(x^1, y) \geq d$ for all $s \in E$, there is a K^δ satisfying

$$\frac{1}{N} \sum_{n=1}^{N} \mathcal{E}_{sx^1y}(R_n) \geq d - \delta \qquad \forall N \geq K^\delta,\ \forall s \in E,$$

where R_n denotes the random variable for the payoff at stage n. Choose a sufficiently large $\tau \in (0,1)$ such that

$$\tau^{K^\delta} \geq 1 - \delta.$$

The strategy x^τ can be interpreted as playing x^1 with probability τ and x^2 with probability $1-\tau$ at each stage, so the last inequality means that x^1 will be played at each K^δ consecutive stages with probability at least $1-\delta$. Hence with probability at least $1-\delta$, the average of the expected payoffs will be at least $d-\delta$ for any K^δ consecutive stages. Therefore, if r denotes the smallest payoff in the game, then for small $\delta > 0$ we have

$$\gamma_s(x^\tau, y) \geq (1-\delta)(d-\delta) + \delta r \geq d - \varepsilon \qquad \forall s \in E,$$

so the proof is complete. □

Now we are ready to prove theorem 3.3.1.

Proof of theorem 3.3.1-(a).
We only show the statement for player 1. First notice that, by the convexity of X^* and by $x \in \text{Relint}(X^*)$, we have $x^\tau_\beta \in \text{Relint}(X^*)$ for all $\beta \in \mathcal{B}$ and $\tau \in (0,1)$.
As there are only finitely many pure stationary strategies, by theorem 2.8.2-(b), it is sufficient to show that, for all $j \in J$, if $\tau \in (0,1)$ and $\beta \in \mathcal{B}$ are large then we have

$$\gamma(x^\tau_\beta, j) \geq v - \varepsilon\, 1_{|S|},$$

where $1_{|S|} = (1, \ldots, 1) \in \mathbb{R}^{|S|}$. Take a $j \in J$ and let $E \subset S$ be an arbitrary ergodic set with respect to (x^τ_β, j). We start with showing that for large $\tau \in (0,1)$ and $\beta \in \mathcal{B}$ we have

$$\gamma_s(x^\tau_\beta, j) \geq v_s - \varepsilon \qquad \forall s \in E. \tag{3.4}$$

Since $x^\tau_\beta \in \text{Relint}(X^*)$, by lemma 3.4.4 we obtain $v_s = v_t := v_E$ for all $s, t \in E$ and $j_s \in J^*_s$ for all $s \in E$. Let $j^*_s := j_s$ for all $s \in E$ and let $j^*_s \in J^*_s$ for all $s \in S \setminus E$; so $j^* \in J^*$. By the definition of x^τ_β and by the properties of $\mathcal{B}$, the set of states E is closed with respect to (x_β, j) for all $\beta \in \mathcal{B}$, so with respect to (x_β, j^*) for all $\beta \in \mathcal{B}$ as well. Thus applying lemma 3.4.3 for Γ^* and using lemma 3.4.2 yield for large $\beta \in \mathcal{B}$ that for all $s \in E$

$$\gamma_s(x_\beta, j) = \gamma_s(x_\beta, j^*) \geq \min_{t \in E} v^*_t - \frac{\varepsilon}{2} \geq \min_{t \in E} v_t - \frac{\varepsilon}{2} = v_E - \frac{\varepsilon}{2}.$$

Now lemma 3.4.5 yields that for large $\tau \in (0,1)$ and for large $\beta \in \mathcal{B}$

$$\gamma_s(x^\tau_\beta, j) \geq v_E - \varepsilon = v_s - \varepsilon \qquad \forall s \in E,$$

which proves (3.4).
Using that $x^\tau_\beta \in X^*$ we have

$$P(x^\tau_\beta, j)\, v \geq v,$$

therefore we obtain inductively for all $n \in \mathbb{N}$

$$P^n(x^\tau_\beta, j)\, v \geq v,$$

which by the definition of $Q(x^\tau_\beta, j)$ yields

$$Q(x^\tau_\beta, j)\, v \geq v.$$

For any $s \in S$, $q_s(t|x^\tau_\beta, j) > 0$ implies that $t \in E$ for some ergodic set E with respect to (x^τ_β, j). Hence by theorem 2.7.1-(c) and (3.4), for large $\tau \in (0,1)$ and $\beta \in \mathcal{B}$ we obtain

$$\begin{aligned}\gamma(x^\tau_\beta, j) &= Q(x^\tau_\beta, j)\,\gamma(x^\tau_\beta, j) \\ &\geq Q(x^\tau_\beta, j)\,(v - \varepsilon\, 1_{|S|}) \\ &= Q(x^\tau_\beta, j)\,v - \varepsilon\, 1_{|S|} \\ &\geq v - \varepsilon\, 1_{|S|},\end{aligned}$$

which completes the proof. □

Proof of theorem 3.3.1-(b).
We only show the statement for player 1. For $n \in \mathbb{N}$, let ε_n and x_n be as in theorem 3.3.1-(b). We will choose an appropriate sequence K_n in $\mathbb{N}$ so that the Markov strategy described in theorem 3.3.1-(b) is optimal. Using the results of Bewley & Kohlberg [1978] (theorem 5.2), for any $n \in \mathbb{N}$, there exists a stage $\bar{K}_n$ such that

$$\frac{1}{N}\sum_{m=1}^{N} \mathcal{E}_{s x_n \sigma}(R_m) \geq v_s - 2\varepsilon_n \qquad \forall N \geq \bar{K}_n,\ \forall s \in S,\ \forall \sigma \in \Sigma, \tag{3.5}$$

where R_m denotes the random variable for the payoff at stage m. Let r denote the smallest payoff in the game minus ε_1 :

$$r := \min_{i_s \in I_s,\, j_s \in J_s,\, s \in S} r_s(i_s, j_s) - \varepsilon_1.$$

Given $\bar{K}_n$, $n \in \mathbb{N}$, choose an arbitrary $K_1 \geq \bar{K}_1$ and choose $K_n \geq \bar{K}_n$, $n \geq 2$, inductively so that

$$\frac{\sum_{l=1}^{n} K_l \cdot (v_s - 2\varepsilon_l) + \bar{K}_{n+1} \cdot r}{\sum_{l=1}^{n} K_l + \bar{K}_{n+1}} \geq v_s - 2\varepsilon_{n-1} \qquad \forall s \in S,\ \forall n \geq 2. \tag{3.6}$$

By the definition of r, inequality (3.6) implies

$$\frac{\sum_{l=1}^{n} K_l \cdot (v_s - 2\varepsilon_l)}{\sum_{l=1}^{n} K_l} \geq v_s - 2\varepsilon_{n-1} \qquad \forall s \in S,\ \forall n \geq 2. \tag{3.7}$$

Let s^1 be an arbitrary initial state and let s^m, $m \geq 2$, denote the random variable for the state at stage m. Let the Markov strategy f be as in theorem 3.3.1-(b). Using that f only prescribes mixed actions in X_s^*, $s \in S$, we have

$$\mathcal{E}_{s^1 f\sigma}(v_{s^m}) \geq v_{s^1} \qquad \forall m \in \mathbb{N},\ \forall \sigma \in \Sigma. \tag{3.8}$$

Let $w_1 := 1$, and let $w_n := \sum_{l=1}^{n-1} K_l + 1$ for all $n \geq 2$. The strategy x_1 is to be played at stages $1, 2, \ldots, K_1$, hence using (3.5)

$$\sum_{m=1}^{K_1} \mathcal{E}_{s^1 f\sigma}(R_m) \geq K_1 \cdot (v_{s^1} - 2\varepsilon_1) \qquad \forall \sigma \in \Sigma; \tag{3.9}$$

while the strategy x_n, $n \geq 2$, is to be played at stages $w_n, \ldots, w_n + K_n - 1$, hence using (3.5) and (3.8) we have

$$\begin{aligned} \sum_{m=w_n}^{w_n+L-1} \mathcal{E}_{s^1 f\sigma}(R_m) &\geq L \cdot \left(\mathcal{E}_{s^1 f\sigma}(v_{s^{w_n}}) - 2\varepsilon_n\right) \\ &\geq L \cdot (v_{s^1} - 2\varepsilon_n) \qquad \forall K_n \geq L \geq \bar{K}_n,\ \forall \sigma \in \Sigma. \end{aligned} \tag{3.10}$$

As a special case, we have for all $n \geq 2$ that

$$\sum_{m=w_n}^{w_n+K_n-1} \mathcal{E}_{s^1 f\sigma}(R_m) \geq K_n \cdot (v_{s^1} - 2\varepsilon_n) \qquad \forall \sigma \in \Sigma. \tag{3.11}$$

Assume first that stage N has the property that for some $n(N) \geq 2$ we have

$$\sum_{l=1}^{n(N)} K_l < N \leq \sum_{l=1}^{n(N)} K_l + \bar{K}_{n(N)+1};$$

which means that at stage N the strategy $x_{n(N)+1}$ is used and it has not yet been used at more than $\bar{K}_{n(N)+1}$ stages. Using (3.9), (3.11), by the definition of r and by (3.6) we have for all $\sigma \in \Sigma$

$$\begin{aligned} \frac{1}{N} \sum_{m=1}^{N} \mathcal{E}_{s^1 f\sigma}(R_m) &= \frac{\sum_{l=1}^{n(N)} \sum_{m=w_l}^{w_l+K_l-1} \mathcal{E}_{s^1 f\sigma}(R_m) + \sum_{m=w_{n(N)+1}}^{N} \mathcal{E}_{s^1 f\sigma}(R_m)}{N} \\ &\geq \frac{\sum_{l=1}^{n(N)} K_l \cdot (v_{s^1} - 2\varepsilon_l) + \left(N - \sum_{l=1}^{n(N)} K_l\right) \cdot r}{N} \end{aligned}$$

$$\geq \frac{\sum_{l=1}^{n(N)} K_l \cdot (v_{s^1} - 2\varepsilon_l) + \bar{K}_{n(N)+1} \cdot r}{\sum_{l=1}^{n(N)} K_l + \bar{K}_{n(N)+1}}$$

$$\geq v_{s^1} - 2\varepsilon_{n(N)-1}. \tag{3.12}$$

Assume now that stage N has the property that for some $n(N) \geq 2$ we have

$$\sum_{l=1}^{n(N)} K_l + \bar{K}_{n(N)+1} < N \leq \sum_{l=1}^{n(N)} K_l + K_{n(N)+1};$$

which means that at stage N the strategy $x_{n(N)+1}$ is used and it has already been used more than $\bar{K}_{n(N)+1}$ stages. Using (3.9), (3.11), (3.10), by (3.7) we have for all $\sigma \in \Sigma$

$$\frac{1}{N} \sum_{m=1}^{N} \mathcal{E}_{s^1 f\sigma}(R_m) = \frac{\sum_{l=1}^{n(N)} \sum_{m=w_l}^{w_l+K_l-1} \mathcal{E}_{s^1 f\sigma}(R_m) + \sum_{m=w_{n(N)+1}}^{N} \mathcal{E}_{s^1 f\sigma}(R_m)}{N}$$

$$\geq \frac{\sum_{l=1}^{n(N)} K_l \cdot (v_{s^1} - 2\varepsilon_l) + \left(N - \sum_{l=1}^{n(N)} K_l\right) \cdot \left(v_{s^1} - 2\varepsilon_{n(N)+1}\right)}{N}$$

$$\geq \frac{\sum_{l=1}^{n(N)} K_l \cdot (v_{s^1} - 2\varepsilon_l)}{\sum_{l=1}^{n(N)} K_l}$$

$$\geq v_{s^1} - 2\varepsilon_{n(N)-1}. \tag{3.13}$$

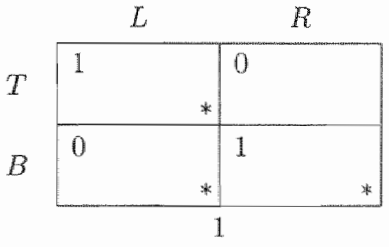
Inequalities (3.12) and (3.13) together imply that

$$\gamma_{s^1}(f, \sigma) = \liminf_{N\to\infty} \frac{1}{N} \sum_{m=1}^{N} \mathcal{E}_{sf\sigma}(R_m) \geq v_{s^1} \qquad \forall \sigma \in \Sigma.$$

Since the initial state s^1 was arbitrary, we have shown that f is optimal in Γ. □

The next example shows that the existence of stationary ε-optimal strategies, for all $\varepsilon > 0$, does not imply the existence of optimal strategies.

Example 3.4.6



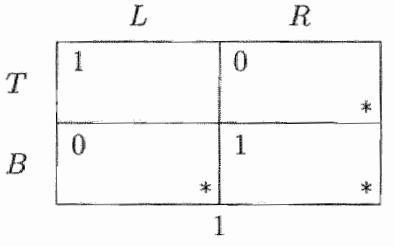
Here the value for initial state 1 is $v_1 = 1$. The stationary strategy for player 1 which, in state 1, prescribes action T with probability $1-\varepsilon$ and action B with probability ε is ε-optimal for small $\varepsilon > 0$. However, we show that player 1 does not have optimal strategies for initial state 1. Take an arbitrary strategy π. We show that player 2 can make sure that player 1's reward is strictly less than 1. Indeed, player 2 has to choose action R as long as π prescribes action T with probability 1 and to play action L at the first stage when π prescribes action B with a positive probability. Then either entry (T, R) is played forever or absorption occurs with payoff zero with a positive probability, thus player 1's reward is strictly less than 1 indeed. $\triangleleft$

3.5 Concluding remarks

Remarks on the restricted game Γ^*. In lemma 3.4.2 we showed that $v_s^* \geq v_s$ for all $s \in S$. In fact, this is the only statement for which we needed the condition that player 1 has an optimal strategy. Therefore if in a zero-sum game $v_s^* \geq v_s$ holds for all $s \in S$, then stationary ε-optimal strategies, $\varepsilon > 0$, and Markov optimal strategies can be constructed exactly as before. It also means that $v_s^* \geq v_s$ for all $s \in S$ holds if and only if player 1 has an optimal strategy.

We also remark that one can find examples in which, although player 1 has an optimal strategy, $v_s^* > v_s$ holds for some state $s \in S$. Such an example is presented next.

Example 3.5.1



The value for initial state 1 is $v_1 = 0$. Indeed, one can easily verify that against the stationary strategy $y = (1-\varepsilon, \varepsilon) \in Y$, for any $\varepsilon \in (0,1)$, player 1 cannot get more than ε. Notice that any strategy of player 1 is optimal. We have $X^* = X$, $Y^* = \{(1,0)\}$, hence $v_1^* = 1 > v_1$. (Obviously, for any absorbing state s in the game $v_s = v_s^*$.) $\lhd$

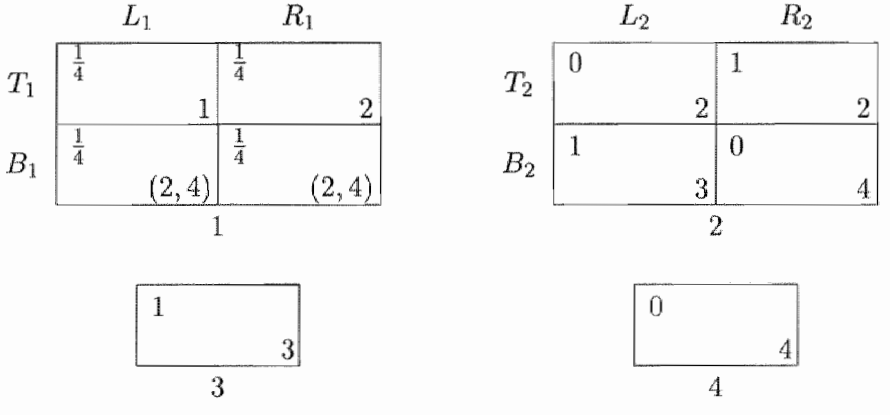
Let E denote an ergodic set with respect to some stationary strategy pair (x,y) in $\text{Relint}(X^*) \times \text{Relint}(Y^*)$. Recall that the value v is a constant v_E on E by lemma 3.4.4. Assume that $v_s^* \geq v_E$ for all $s \in E$ (as we discussed above, this assumption follows if player 1 has an optimal strategy). We will now show that there must exist a state $s \in E$ such that $v_s^* = v_E$. To see this one can argue as follows. Suppose to the contrary that $v_s^* \geq v_E + \mu$ for all $s \in E$, where $\mu > 0$. Let $x_\beta^\tau \in \text{Relint}(X^*)$ be defined as in theorem 3.3.1-(a). Then one can show, by using similar arguments as in the proof of theorem 3.3.1-(a), that for large τ and β we have

$$\gamma_s(x_\beta^\tau, j) \geq \min_{t \in E} v_t^* - \frac{\mu}{2} \geq v_E + \frac{\mu}{2} \qquad \forall s \in E,\ \forall j \in J^*. \tag{3.14}$$

Let player 1 play the strategy π^δ, $\delta > 0$, which prescribes to play as follows: play x_β^τ as long as player 2 chooses actions in J_s^*, $s \in E$, and start playing a δ-optimal strategy as soon as player 2 chooses an action in $J_s \setminus J_s^*$ in some state $s \in E$. Note that if player 2 always chooses actions in J_s^*, $s \in E$, then (3.14) assures that the reward is at least $v_E + \frac{\mu}{2}$ (recall theorem 2.8.2-(b)). On the other hand, if player 2 chooses an action in $J_s \setminus J_s^*$ in some state $s \in E$, then one can show that $x_{\beta s}^\tau \in \text{Relint}(X_s^*)$ yields that the original value v increases in expectation by at least some $\nu > 0$; so if $\delta \in (0, \frac{\nu}{2})$, by the definition of π^δ, the reward is at least $v_E + \frac{\nu}{2}$ in this case. Therefore π^δ, with $\delta \in (0, \frac{\nu}{2})$, guarantees a reward at least $v_E + \frac{1}{2}\min(\mu, \nu) > v_E$, which contradicts the definition of the value. So $v_s^* = v_E$ must hold for some state $s \in E$.

Optimal strategies for particular initial states. We briefly discuss a generalization of the results, which concerns strategies that are only optimal for particular initial states. Let $\tilde{S}$ denote the set of states for which player 1 has an optimal strategy. First note that, in each stochastic game, there exists at least one initial state for which player 1 has optimal strategies (cf. Thuijsman & Vrieze [1993]), so the set $\tilde{S}$ is always nonempty. Using similar techniques as before, one can show that, for any $\varepsilon > 0$, player 1 has a strategy ξ^ε which for all initial states $\alpha \in \tilde{S}$ satisfies: (a) ξ^ε is ε-optimal, (b) ξ^ε is stationary as long as the play is in $\tilde{S}$, (c) there exist stationary best replies of player 2 against ξ^ε, (d) the probability of ever leaving $\tilde{S}$ is zero with respect to $(\xi^\varepsilon, \sigma)$ and initial state α, if σ is a best reply. The weakness of this result is due to the fact that stationary strategies are not effective in states outside $\tilde{S}$, so player 1 may have to start playing a history dependent δ-optimal strategy if the play leaves $\tilde{S}$, for some $\delta > 0$. One can also show that player 1 has analogous "almost Markov" optimal strategies for all initial states $\alpha \in \tilde{S}$.

Example 3.5.2



This example clarifies the existence of such "almost stationary" ε-optimal strategies for initial states in $\tilde{S}$. The two "mixed" transition vectors in entries (B_1, L_1) and (B_1, R_1) lead to state 2 with probability $\frac{1}{2}$ and to state 4 with probability $\frac{1}{2}$. Notice that if the initial state is state 2 then this game reduces to the Big Match (cf. example 2.9.6 and lemma 2.9.7). So here the value is $v = (\frac{1}{4}, \frac{1}{2}, 1, 0)$ and player 1 has no optimal strategy for initial state 2, so $\tilde{S} = \{1, 3, 4\}$. Since initial states $3, 4 \in \tilde{S}$ are trivial, we assume that the initial state is state $1 \in \tilde{S}$.

Consider the strategy ξ for player 1 which prescribes to play action T_1 as long as the play is in state 1 and as soon as the play visits state 2 then prescribes to start playing a history dependent $\frac{1}{8}$-optimal strategy. This strategy ξ is optimal and clearly satisfies properties (a), (b), (c), and (d). Note that switching to a history dependent strategy when entering state 2 is crucial, because by stationary strategies player 1 could only guarantee 0 for initial state 2 (cf. lemma 2.9.7-(e)). Note also that even though $(0,1) \in X_1^*$, player 1 should not choose action B_1, because it would violate property (d).

Subgame optimality. Note that the Markov strategy f constructed in theorem 3.3.1-(b) is "subgame optimal"; namely the strategy $f[h]$ is optimal for any finite history $h \in H$. (The stationary ε-optimal strategies, $\varepsilon > 0$, in theorem 3.3.1-(a) are obviously "subgame ε-optimal", as $x = x[h]$ for any stationary strategy $x \in X$ and for any finite history $h \in H$).

Chapter 4

Improving and non-improving strategies

4.1 Introduction

In zero-sum stochastic games the players have completely opposite interests, so it is natural to evaluate a strategy of a player by the reward it guarantees against any strategy of the opponent. So as in definition 2.9.1, for strategies $\pi \in \Pi$ and $\sigma \in \Sigma$ let

$$\underline{v}_s(\pi) := \inf_{\sigma' \in \Sigma} \gamma_s(\pi, \sigma') \quad \forall s \in S, \qquad \underline{v}(\pi) := (\underline{v}_s(\pi))_{s \in S},$$

$$\overline{v}_s(\sigma) := \sup_{\pi' \in \Pi} \gamma_s(\pi', \sigma) \quad \forall s \in S, \qquad \overline{v}(\sigma) := (\overline{v}_s(\sigma))_{s \in S}.$$

These evaluations enable us to compare strategies as follows.

Definition 4.1.1

(a) *A strategy π^1 is called ε-better than π^2, where $\varepsilon \geq 0$, if*

$$\underline{v}_s(\pi^1) \geq \underline{v}_s(\pi^2) - \varepsilon \qquad \forall s \in S.$$

0-better strategies are simply called better. A similar definition of ε-betterness holds for strategies of player 2.

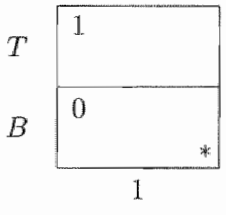
(b) *A strategy π is called non-improving if*

$$\underline{v}_s(\pi) \geq \underline{v}_s(\pi[h]) \qquad \forall s \in S$$

holds for any history $h \in H$; otherwise π is called improving. Non-improving strategies and improving strategies for player 2 are similarly defined.

Intuitively, a non-improving strategy cannot guarantee a larger reward conditional on any past history than initially. On the other hand, improving strategies may become better during the play than initially.
For example, all stationary strategies are clearly non-improving strategies, because $x = x[h]$ for any history $h \in H$. In the following simple example we show an instance of an improving strategy.

Example 4.1.2



Consider the Markov strategy f for player 1 which prescribes to play action T with probability $1/2$ and action B with probability $1/2$ at stage 1, and if the play does not absorb then to play action T at all further stages. Clearly, f yields reward $1/2$, hence we obtain $\underline{v}_1(f) = 1/2$. However, if h denotes the history up to stage 1 when player 1 chooses action T at stage 1, then the strategy $f[h]$ prescribes action T for each stage, hence $\underline{v}_1(f[h]) = 1$. Thus $\underline{v}_1(f) < \underline{v}_1(f[h])$, which means that f is improving. ◁

In this chapter, which is based on Flesch et. al. [1998,IV], we derive several results on the use of non-improving and improving strategies. The main results are summarized as follows.

Main Theorem 4 *In any zero-sum stochastic game, for any non-improving strategy, there exists an ε-better stationary strategy, for any $\varepsilon > 0$, and there exists a better Markov strategy as well.*

The above theorem will follow from theorem 4.3.1 below. It says that, surprisingly, non-improving strategies are not more effective than stationary strategies or Markov strategies. This also means that, instead of using a complex history dependent non-improving strategy, the player could also use a simple stationary strategy which guarantees at least the same reward up to some arbitrarily small $\varepsilon > 0$, or he could even achieve the same reward by employing a Markov strategy.
Notice that optimal strategies are always non-improving, since they guarantee the value and no higher reward can be guaranteed, by the definition of the value. In light of this observation, the above result can be seen as a generalization of Main Theorem 3 in chapter 3.
The above theorem and Main Theorem 3 together have the following corollary, which shows the insufficiency of the class of non-improving strategies as well as the indispensability of improving strategies for achieving ε-optimality, for small $\varepsilon > 0$.

Corollary 4.1.3 *In any zero-sum stochastic game, if a player has no stationary ε-optimal strategies for all $\varepsilon > 0$, then he has no optimal strategies either and all his ε-optimal strategies, with small $\varepsilon > 0$, are improving.*

The next example provides an illustration.

Example 4.1.4

	L	R
T	0	1
B	1 *	0 *

1

This example is the Big Match in example 2.9.6. In lemma 2.9.7-(a),(c) we discussed that the value for initial state 1 is $v_1 = 1/2$ and player 1 has no stationary ε-optimal strategies for small $\varepsilon > 0$. By the above corollary, we should have that player 1 has no optimal strategies and all his ε-optimal strategies, with small $\varepsilon > 0$, are improving.
Indeed, in view of lemma 2.9.7-(d), player 1 has no optimal strategies. Now consider the history dependent ε-optimal strategies π^N in lemma 2.9.7-(c). It is easy to verify that the strategies π^N are improving, since for the history $h = (1, T, R)$ we have $\pi^N[h] = \pi^{N+1}$. $\lhd$

4.2 Preliminaries

The definitions will be very similar to those in chapter 3. Let π denote a fixed non-improving strategy and let

$$a := \underline{v}(\pi).$$

For $x \in X$, $y \in Y$, $s \in S$ let

$$A_s(x_s, y_s) := \sum_{t \in S} p_s(t|x_s, y_s)\, a_t, \quad A(x,y) := (A_s(x_s, y_s))_{s \in S}.$$

Let

$$\tilde{X}_s := \{x_s \in X_s |\, A_s(x_s, y_s) \geq a_s \;\; \forall y_s \in Y_s\} \quad \forall s \in S, \quad \tilde{X} := \times_{s \in S} \tilde{X}_s,$$

so $\tilde{X}_s$ is the set of mixed actions of player 1 in state s which assure that after transition a will not decrease in expectation.

Lemma 4.2.1 *The sets $\tilde{X}_s$, $s \in S$, are nonempty polytopes.*

Proof. Let $s \in S$. One can verify that the linearity of A_s in both components implies that the set $\tilde{X}_s$ is a polytope.
Now we prove that $\tilde{X}_s$ is nonempty by showing that $\pi_s \in \tilde{X}_s$. (Recall that π_s denotes the mixed action prescribed by π for stage 1 if the initial state is state s). By the definition of $\underline{v}_s(\pi)$

$$\underline{v}_s(\pi) = \min_{y_s \in Y_s} \sum_{t \in S} \sum_{i_s \in I_s, j_s \in J_s} [\pi_s(i_s) \cdot y_s(j_s) \cdot p_s(t|i_s, j_s)] \cdot \underline{v}_t(\pi[s, i_s, j_s]),$$

hence using the definition of a and the non-improvingness of π we have

$$\begin{aligned}
a_s &= \underline{v}_s(\pi) \\
&= \min_{y_s \in Y_s} \sum_{t \in S} \sum_{i_s \in I_s, j_s \in J_s} [\pi_s(i_s) \cdot y_s(j_s) \cdot p_s(t|i_s, j_s)] \cdot \underline{v}_t(\pi[s, i_s, j_s]) \\
&\leq \min_{y_s \in Y_s} \sum_{t \in S} \sum_{i_s \in I_s, j_s \in J_s} [\pi_s(i_s) \cdot y_s(j_s) \cdot p_s(t|i_s, j_s)] \cdot \underline{v}_t(\pi) \\
&= \min_{y_s \in Y_s} \sum_{t \in S} p_s(t|\pi_s, y_s) \cdot a_t \\
&= \min_{y_s \in Y_s} A_s(\pi_s, y_s),
\end{aligned}$$

so the proof is complete. □

As in chapter 3, if Z is a polytope then Relint(Z) denotes the relative interior of the polytope Z, which is defined as the set of points in Z that can be written as a convex combination of all the extreme points of Z with only strictly positive coefficients.
The following technical lemma is needed later for the construction of a restricted game. Here, on condition that player 1 uses a strategy $x \in \text{Relint}(\tilde{X})$, we are looking for the largest set S' of states which can be made recurrent and the largest sets Y'_s, $s \in S'$, of mixed actions which keep all the states in S' recurrent.

Lemma 4.2.2 *There exist a nonempty $S' \subset S$ and a nonempty $Y' = \times_{s \in S'} Y'_s$, where $Y'_s \subset Y_s$ are polytopes for all $s \in S'$, such that for any $x \in$ Relint$(\tilde{X})$*

(a) *for any $y \in Y$, if $s \in S$ is recurrent with respect to (x, y) then $s \in S'$ and $y_s \in Y'_s$;*

(b) *for any $y \in Y$ with $y_s \in$ Relint(Y'_s) for all $s \in S'$, all states $s \in S'$ are recurrent with respect to (x, y).*

Proof. Take an arbitrary $x \in \text{Relint}(\tilde{X})$. For $j \in J$, let $R(j)$ denote the set of recurrent states with respect to (x, j). Now let

$$S' := \cup_{j \in J} R(j).$$

For $s \in S'$ let

$$J'_s := \{j_s \in J_s | \, \exists \bar{j} \in J : \; \bar{j}_s = j_s, \; s \in R(\bar{j})\},$$

$$Y'_s := \text{conv}\left\{J'_s\right\}, \quad Y' := \times_{s \in S'} Y'_s,$$

where conv stands for the convex hull of a set. Note that these sets are independent of the choice of $x \in \text{Relint}(\tilde{X})$, because all $x \in \text{Relint}(\tilde{X})$ put positive probabilities on the same actions in any state. It is not hard to check that S' and Y' satisfy the required properties. □

4.3 The construction

Recall that we have fixed a non-improving strategy π for player 1. Let $\tilde{X}$ be as above, let S' and Y' be as in lemma 4.2.2, and let $X' := \times_{s \in S'} \tilde{X}_s$. In view of lemma 4.2.2, any pair of mixed actions in $X'_s \times Y'_s$ in any state $s \in S'$ only

induces transitions to states in S'. Hence, we may define a restricted stochastic game Γ', as described in section 3.2, in the following way. Let Γ' be the game, derived from Γ, where the state space is S' and the players are restricted to using mixed actions in X'_s and Y'_s if the play is in any state $s \in S'$. The respective stationary strategy spaces in Γ' for the players are X' and Y'.

By the finiteness of the state and action spaces, there exists a countable subset of discount factors $\mathcal{B} \subset (0,1)$ with the properties that 1 is a limit point of $\mathcal{B}$ and there are stationary β-discounted optimal strategies $x_\beta \in X'$ in the restricted game Γ' such that the sets

$$\{i_s \in I_s | \, x_{\beta s}(i_s) > 0\}, \qquad s \in S,$$

are independent of $\beta \in \mathcal{B}$. In this chapter, each time that we are dealing with discount factors, discounted optimal strategies, or with limits when the discount factors converge to 1, we will have such a subset of discount factors $\mathcal{B}$ in mind.

Theorem 4.3.1 *Let $\pi \in \Pi$ be a non-improving strategy in a zero-sum stochastic game. Given the strategy π, let S', X', Y', and the restricted game Γ' be as above.*

(a) *For any $\beta \in \mathcal{B}$, let $x_\beta \in X'$ be a β-discounted optimal strategy in the restricted game Γ' and let $x \in Relint(\tilde{X})$. For $\beta \in \mathcal{B}$ and $\tau \in (0,1)$, let the stationary strategy $x^\tau_\beta \in \tilde{X}$ be given by*

$$x^\tau_{\beta s} := \begin{cases} \tau \cdot x_{\beta s} + (1-\tau) \cdot x_s & \text{if } s \in S' \\ x_s & \text{if } s \in S \setminus S' \end{cases} \qquad \forall s \in S.$$

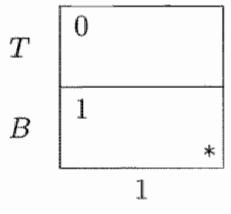
Then, for any $\varepsilon > 0$, if $\beta \in \mathcal{B}$, $\tau \in (0,1)$ are sufficiently large then the stationary strategy $x^\tau_\beta \in \tilde{X}$ is ε-better than π in the game Γ.

(b) *Let ε_n, $n \in \mathbb{N}$, be a strictly monotonously decreasing sequence which converges to 0. Let the stationary strategy $x_n \in X'$ be ε_n-better than π for all $n \in \mathbb{N}$. Then there exists a sequence K_n in $\mathbb{N}$ such that the Markov strategy f which prescribes to play x_1 for the first K_1 stages, then to play x_2 for the next K_2 stages, and so on, is better than π in the game Γ.*

A similar statement holds for player 2 as well.

Since optimal strategies are always non-improving, the above theorem can be seen as a generalization of theorem 3.3.1 (even though the restricted game was somewhat differently defined in chapter 3). So one may read example 3.3.2 for an illustration; just instead of an optimal strategy one has to think of a non-improving strategy which guarantees the value (so a equals the value v in this case). We will now explain with the help of the following example why the proof of theorem 3.3.1 does not apply directly.

Example 4.3.2



Consider the stationary strategy x which prescribes action T in state 1 with probability 1. Clearly, this strategy is non-improving and we have $a_1 = 0$. Notice that x is not effective in state 1, since player 1 would be better off by choosing action B. Formally, it means that there exists a mixed action $\bar{x}_1 \in X_1$ in state 1 so that we have

$$A_1(\bar{x}_1, y_1) > a_1 \qquad \forall y_1 \in Y_1.$$

Generally, such states cannot be dealt with in the same way as in the proof of theorem 3.3.1, since a version of lemma 3.2.4 would not hold here.
Nevertheless, in states in S' the construction remains almost the same. Here S' only consists of the trivial absorbing state. $\lhd$

4.4 The proof

In this section we provide a proof of theorem 4.3.1, which will go along similar lines as the proof of theorem 3.3.1 in section 3.4. Recall that we have fixed a non-improving strategy π. In the restricted game Γ', let H' denote the set of finite histories, Π' and Σ' the sets of history dependent strategies, γ' the average reward, and v'_β the β-discounted value for all $\beta \in (0, 1)$.

Let $v' := \lim_{\beta \uparrow 1} v'_\beta$ (as discussed in section 3.2, the limit must exist). Moreover, let

$$\bar{\Pi} := \left\{ \pi \in \Pi | \, \pi_s(h) \in X'_s \quad \text{for all } s \in S' \text{ and } h \in H' \right\}$$

$$\bar{\Sigma} := \left\{ \sigma \in \Sigma | \, \sigma_s(h) \in Y'_s \quad \text{for all } s \in S' \text{ and } h \in H' \right\};$$

so $\bar{\Pi}$ and $\bar{\Sigma}$ are the sets of strategies in the original game Γ which behave as strategies in Π' and Σ' as long as the play is in the restricted game Γ'.
By using the definition of $\tilde{X}$, the following lemma follows analogously to the first part of lemma 3.4.4.

Lemma 4.4.1 *Let $x \in \tilde{X}$ and $y \in Y$. Suppose that E is an ergodic set with respect to (x, y). Then $a_s = a_t$ for all $s, t \in E$.*

Next, we show an important property of the sets Y'_s, $s \in S'$.

Lemma 4.4.2 *Let $s \in S'$. Then $A_s(x_s, y_s) = a_s$ for all $x_s \in X'_s$ and $y_s \in Y'_s$.*

Proof. Take arbitrary $x_s \in X'_s$ and $y_s \in Y'_s$. Let $\bar{x} \in \text{Relint}(\tilde{X})$ and $\bar{y} \in Y$ with $\bar{y}_t \in \text{Relint}(Y'_t)$ for all $t \in S'$. In view of lemma 4.2.2-(b), state s belongs to an ergotic set E with regard to $(\bar{x}, \bar{y})$, hence by lemma 4.4.1, we obtain $a_s = a_t$ for all $t \in E$. As $p_s(t|\bar{x}_s, \bar{y}_s) > 0$ implies $t \in E$, it must also hold that $p_s(t|x_s, y_s) > 0$ implies $t \in E$. Therefore the proof is complete. □

The next lemma is similar to lemma 3.4.1. It says that, as long as player 2 plays in the restricted game Γ', the non-improving strategy π behaves as a strategy in $\bar{\Pi}$.

Lemma 4.4.3 *Let $s \in S'$ and let*

$$U_s := \{(h, t) \in H_s \times S | \, \exists \sigma \in \bar{\Sigma} : \quad \mathcal{P}_{s\pi\sigma}(h) > 0 \quad \textit{and} \quad \mathcal{P}_{s\pi\sigma}(t|h) > 0\},$$

where $\mathcal{P}_{s\pi\sigma}(t|h)$ is the probability that, with respect to (π, σ), state t becomes the new state after history h. Then $h \in H'_s$, $t \in S'$, $\pi_t(h) \in X'_t$ for all $(h, t) \in U_s$.

Proof. Suppose the opposite. Then, there is a shortest history $\bar{h}^n \in H_s$, say up to stage n, with the following properties: there exist $\sigma \in \bar{\Sigma}$ and a state t such that

$$\mathcal{P}_{s\pi\sigma}(\bar{h}^n) > 0 \quad \text{and} \quad \mathcal{P}_{s\pi\sigma}(t|\bar{h}^n) > 0,$$

and

$$\bar{h}^n \in H_s \setminus H'_s \quad \text{or} \quad t \in S \setminus S' \quad \text{or} \quad \pi_t(\bar{h}^n) \in X_t \setminus X^*_t.$$

Similarly to the proof of lemma 3.4.1, one can show that

$$\bar{h}^n \in H'_s \quad \text{and} \quad t \in S'$$

necessarily hold. Therefore we must have $\pi_t(\bar{h}^n) \in X_t \setminus X'_t$. Hence, there exists a $\bar{y}_t \in Y_t$ such that

$$\tau := a_t - A_t(\pi_t(\bar{h}^n), \bar{y}_t) > 0.$$

For any present state $z \in S'$ and past history $h \in H'$, we define a mixed action $\phi_z(h) \in Y_z$ as follows: if $\pi_z(h) \in X'_z$ then let $\phi_z(h) \in Y'_z$; while if $\pi_z(h) \in X_z \setminus X'_z$ then let $\phi_z(h) \in Y_z$ such that $A_z(\pi_z(h), \phi_z(h)) \leq a_z$. By lemma 4.4.2, we have that

$$A_z(\pi_z(h), \phi_z(h)) \leq a_z \qquad \forall z \in S',\ \forall h \in H'. \tag{4.1}$$

For completeness, take an arbitrary $\phi_z(h) \in Y_z$ if $z \in S \setminus S'$ or $h \in H \setminus H'$. Let

$$\delta \in \left(0, \mathcal{P}_{s\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s\pi\sigma}(t|\bar{h}^n) \cdot \tau\right).$$

Let $s^1 := s$, and let s^m, $m \geq 2$, denote the random variable for the state at stage m, and let θ^m denote random variable for the history up to stage $m \in \mathbb{N}$. Let $\sigma^\delta \in \Sigma$ be the strategy that prescribes to play as follows: play σ during the first n stages; at stage $n+1$, if $\theta^n = \bar{h}^n$ and $s^{n+1} = t$ then play $\bar{y}_t$ while if $\theta^n \neq \bar{h}^n$ or $s^{n+1} \neq t$ then play the mixed action $\phi_{s^{n+1}}(\theta^n)$; and finally, play a δ-best reply against $\pi[\theta^{n+1}]$ from stage $n+2$ on. Note that

$$\mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) = \mathcal{P}_{s^1\pi\sigma^\delta}(\bar{h}^n) > 0, \qquad \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) = \mathcal{P}_{s^1\pi\sigma^\delta}(t|\bar{h}^n) > 0.$$

Since we have chosen a shortest history $\bar{h}^n$ with its above specified properties, the play up to stage n must take place in the restricted game Γ'. Hence, lemma 4.4.2 implies

$$\mathcal{E}_{s^1\pi\sigma^\delta}(a_{s^{n+1}}) = a_{s^1}.$$

In view of lemma 4.2.2 we also have $s^{n+1} \in S'$. By the choices of the used mixed actions at stage $n+1$ and by (4.1)

$$\mathcal{E}_{s^1\pi\sigma^\delta}(a_{s^{n+2}}) \leq \mathcal{E}_{s^1\pi\sigma^\delta}(a_{s^{n+1}}) - \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) \cdot \tau.$$

For $h^{n+1} \in H_s^{n+1}$ and $z \in S$ let

$$\mathcal{P}_{s^1\pi\sigma}(h^{n+1}, z) := \mathcal{P}_{s^1\pi\sigma}(h^{n+1}) \cdot \mathcal{P}_{s^1\pi\sigma}(z|h^{n+1}).$$

Since player 2 plays a δ-best reply from stage $n+2$ on and π is non-improving, the choice of δ yields

$$\begin{aligned}
\gamma_{s^1}(\pi, \sigma^\delta) &\leq \sum_{\substack{h^{n+1} \in H_s^{n+1} \\ z \in S}} \mathcal{P}_{s^1\pi\sigma}(h^{n+1}, z) \cdot \gamma_z(\pi[h^{n+1}], \sigma^\delta[h^{n+1}]) \\
&\leq \sum_{\substack{h^{n+1} \in H_s^{n+1} \\ z \in S}} \mathcal{P}_{s^1\pi\sigma}(h^{n+1}, z) \cdot \left(\underline{v}_z(\pi[h^{n+1}]) + \delta\right) \\
&\leq \sum_{\substack{h^{n+1} \in H_s^{n+1} \\ z \in S}} \mathcal{P}_{s^1\pi\sigma}(h^{n+1}, z) \cdot (\underline{v}_z(\pi) + \delta) \\
&\leq \sum_{\substack{h^{n+1} \in H_s^{n+1} \\ z \in S}} \mathcal{P}_{s^1\pi\sigma}(h^{n+1}, z) \cdot (a_z + \delta) \\
&= \mathcal{E}_{s^1\pi\sigma^\delta}(a_{s^{n+2}}) + \delta \\
&\leq \mathcal{E}_{s^1\pi\sigma^\delta}(a_{s^{n+1}}) - \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) \cdot \tau + \delta \\
&= a_{s^1} - \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) \cdot \tau + \delta \\
&< a_{s^1},
\end{aligned}$$

which contradicts the definition of a. □

Using the previous lemma, the next result follows similarly to lemma 3.4.2.

Lemma 4.4.4 *For any $s \in S$*

$$v_s \leq \sup_{\pi' \in \Pi'} \inf_{\sigma' \in \Sigma'} \gamma'_s(\pi', \sigma') \leq v'_s.$$

We will now turn to the proof of theorem 4.3.1.

Proof of theorem 4.3.1:
We will only show the statement for player 1. It is sufficient to focus on part (a), because the proof of (b) is the same as the proof of theorem 3.3.1-(b), just a has to be used instead of v.
We will now show part (a). Since there are only finitely many pure stationary strategies, by theorem 2.8.2-(b), it suffices to show that, for all $j \in J$, if $\tau \in (0,1)$ and $\beta \in \mathcal{B}$ are large then

$$\gamma(x_\beta^\tau, j) \geq a - \varepsilon\, 1_{|S|},$$

where $1_{|S|} = (1, \ldots, 1) \in \mathbb{R}^{|S|}$.
Take a $j \in J$ and let $E \subset S$ be an arbitrary ergodic set with respect to (x_β^τ, j). Notice that, by the definitions, we have $x_\beta^\tau \in \text{Relint}(\tilde{X})$ for all $\beta \in \mathcal{B}$ and $\tau \in (0,1)$. Therefore lemma 4.2.2-(a) implies that $E \subset S'$ and $j_s \in Y_s'$ for all $s \in E$, so the play takes place in the restricted game Γ' in states in E. Based on the above lemmas, one can show, along similar lines as (3.4) in the proof of theorem 3.3.1-(a), that for large $\tau \in (0,1), \beta \in \mathcal{B}$ we have

$$\gamma_s(x_\beta^\tau, j) \geq a_s - \varepsilon \qquad \forall s \in E.$$

Now the rest of the proof is the same as the proof of theorem 3.3.1-(a), just a has to be used instead of v, so the proof is complete. □

Chapter 5

Markov strategies are better than stationary strategies

5.1 Introduction

In zero-sum stochastic games, as before, we evaluate strategies of player 1 by the highest rewards they guarantee against any strategy of his opponent, player 2. For the sake of simplicity, we only focus on strategies of player 1 here. So as in definition 2.9.1, for $\pi \in \Pi$ let

$$\underline{v}_s(\pi) := \inf_{\sigma \in \Sigma} \gamma_s(\pi, \sigma) \quad \forall s \in S, \qquad \underline{v}(\pi) := (\underline{v}_s(\pi))_{s \in S}.$$

For an initial state $s \in S$, the highest reward that can be guaranteed by stationary strategies is called the stationary utility, denoted by $\mathcal{A}_s$, and the highest reward that can be guaranteed by Markov strategies is called the Markov utility, denoted by $\mathcal{B}_s$. Formally,

$$\mathcal{A}_s := \sup_{x \in X} \underline{v}_s(x), \quad \mathcal{B}_s := \sup_{f \in F} \underline{v}_s(f), \quad \mathcal{A} := (\mathcal{A}_s)_{s \in S}, \quad \mathcal{B} := (\mathcal{B}_s)_{s \in S}.$$

The fact that all stationary strategies are Markov strategies as well, and the definition of the value yield

$$\mathcal{A}_s \leq \mathcal{B}_s \leq v_s \qquad \forall s \in S. \tag{5.1}$$

Although the class of Markov strategies is richer than the class of stationary strategies, so far no substantial difference has been found in the use of stationary and Markov strategies in zero-sum stochastic games (with finite state and action spaces). Most classes of stochastic games, examined so far, have the

property that both players have stationary ε-optimal strategies for all $\varepsilon > 0$ (cf. theorem 2.11.2). Thus, in view of (5.1), in those classes Markov strategies do not yield higher rewards than stationary strategies. The only class in which stationary ε-optimal strategies are not available is the class of repeated games with absorbing states. Later we provide a direct proof that the equality $\mathcal{A} = \mathcal{B}$ holds for this class as well (the very same result also follows from a more general result in Coulomb [1992]).
The goal of this chapter, which is based on Flesch et al. [1997,II], is to explore the way how Markov strategies can be more effective than stationary strategies, as well as to find sufficient conditions under which these two classes of strategies perform equally well.

The main results can be summarized as follows.

Main Theorem 5

(a) *There are zero-sum stochastic games in which $\mathcal{A}_s < \mathcal{B}_s$ holds for particular initial states $s \in S$.*

(b) *In every zero-sum stochastic game, there exists an initial state $s \in S$ for which $\mathcal{A}_s = \mathcal{B}_s$.*

(c) *We have $\mathcal{A} = \mathcal{B}$ in*

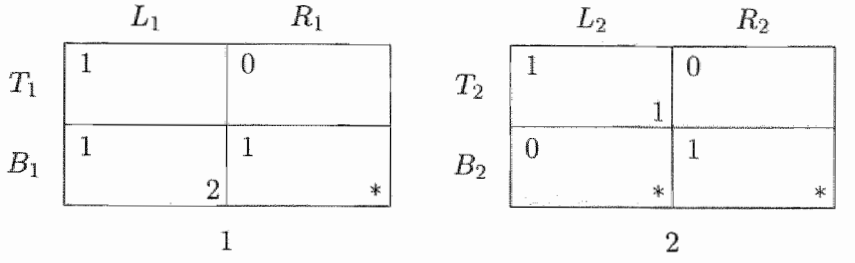
- *repeated games with absorbing states;*
- *games where for all $s, t \in S$ either $\mathcal{A}_s = \mathcal{A}_t$ or $\mathcal{B}_s = \mathcal{B}_t$ holds; in particular, in games with constant $\mathcal{A}$ or $\mathcal{B}$;*
- *games in which player 1 has an optimal strategy;*
- *games in which player 1 has a best Markov strategy, namely a Markov strategy $f \in F$ satisfying $\underline{v}(f) \geq \underline{v}(f')$ for all Markov strategies $f' \in F$.*

In the above theorem, part (a) will follow from example 5.2.1 and theorem 5.2.2, part (b) from theorem 5.3.1, while part (c) from theorems 5.3.2, 5.3.6, and 5.3.7.

5.2 An example where $\mathcal{A} < \mathcal{B}$ for some initial states

This section is devoted to the following example demonstrating that $\mathcal{B}$ may be strictly larger than $\mathcal{A}$ for some initial states.

Example 5.2.1 *Game* Γ :



The main result of this section is the next theorem, which will follow from lemmas 5.2.3 and 5.2.11 below.

Theorem 5.2.2 *In the game* Γ *we have* $0 = \mathcal{A}_t < \mathcal{B}_t = 1 = v_t$ *for initial states* $t = 1, 2$.

This theorem states that, for initial states 1 and 2 in the game Γ, player 1 can get at most 0 by using stationary strategies, although he can get as close to 1 as he likes if he applies Markov strategies.

Since there are two actions for each player in states 1 and 2, we may represent each mixed action in state 1 and in state 2 by the probability assigned to the first action, which lets the stationary and Markov strategy spaces be

$$X = Y = [0,1] \times [0,1], \qquad F = G = \times_{n \in \mathbb{N}} ([0,1] \times [0,1]) .$$

First we intuitively discuss the main steps of the proof. We will start with an easy proof that $\mathcal{A}_t = 0$ for initial states $t = 1, 2$ (cf. lemma 5.2.3). Since the largest payoff in the game is 1, in view of (5.1), it remains to show that $\mathcal{B}_t = 1$ for initial states $t = 1, 2$. However, for this step we need to analyze the game in detail. We define a Markov strategy f^K for player 1 where $K \in \mathbb{N}$: let

$$u^K(n) := \sqrt[K]{\frac{n}{n+1}} \quad \text{for all } n \in \mathbb{N}, \qquad f^K := (u^K(n), u^K(n))_{n \in \mathbb{N}} \in F.$$

Observe that the Markov strategy f^K is symmetric in the sense that the prescribed mixed actions in state 1 and state 2 are the same for any stage. Note that the sequence $u^K(n)$ converges to 1 as n tends to infinity, so f^K assigns less and less probabilities to actions B_1 and B_2.
For initial states 1 and 2, we will show that, for all $\varepsilon > 0$, if $K \in \mathbb{N}$ is large then player 1 can guarantee a reward at least $1 - \varepsilon$ by playing the Markov strategy f^K.

Now the question is how player 2 can reply to the strategy f^K. Intuitively, player 2 has two hopes to decrease player 1's reward. The first one is achieving absorption in entry (B_2, L_2) with payoff 0. Player 2's best candidate would be playing actions L_1 and L_2 whenever the play is in state 1 or in state 2. But then whenever the play is in state 2, a transition occurs to state 1 with a large probability, and from state 1 it takes a long time, and for large stages even a longer and longer time, until the play comes back to state 2 again. So using that the strategy f^K assigns less and less probabilities to B_2, the probability of absorption in entry (B_2, L_2) turns out to be small (cf. lemma 5.2.8). On the other hand, using that the payoffs in entries (T_1, R_1) and (T_2, R_2) equal 0, player 2 could try to play actions R_1 and R_2 "often enough" and hope that the play will not absorb. But in this case it will appear that the play will eventually absorb with probability 1 (cf. lemma 5.2.9), and then the zero payoffs in entries (T_1, R_1), (T_2, R_2) will have no influence on the reward.
First we show that, by playing stationary strategies, player 1 can get at most 0 for initial states 1 and 2.

Lemma 5.2.3 *$\mathcal{A}_t = 0$ for initial states $t = 1, 2$ in the game Γ.*

Proof. For each strategy $x = (x_1, x_2)$ we define a strategy $y^x = (y_1^x, y_2^x)$ for player 2: let

$$y_1^x := \begin{cases} 1 & \text{if } x_1 < 1 \\ 0 & \text{if } x_1 = 1 \end{cases}, \qquad y_2^x := \begin{cases} 1 & \text{if } x_2 < 1 \\ 0 & \text{if } x_2 = 1 \end{cases}.$$

Notice that, for $t = 1, 2$, we have $\gamma_t(x, y^x) = 0$ for all $x \in X$, so

$$\mathcal{A}_t = \sup_{x \in X} \underline{v}_t(x) = \sup_{x \in X} \inf_{\sigma \in \Sigma} \gamma_t(x, \sigma) \leq \sup_{x \in X} \gamma_t(x, y^x) = 0 \qquad \forall t = 1, 2.$$

Since the smallest payoff in the game is 0, the proof is complete. □

For the analysis of f^K, defined above, we need two important properties of the speed of convergence when $u^K(n)$ tends to 1 as n goes to infinity. The first property says that the convergence is fast in the sense that, intuitively, for any $\varepsilon > 0$, if $K \in \mathbb{N}$ is sufficiently large then the probability of ever playing action B_1 or action B_2 at stages 2^{n-1}, $n \in \mathbb{N}$, is at most $\varepsilon/2$. However, the second property tells us that, in a "dense" set of stages, one of the bottom actions B_1 and B_2 will eventually be chosen, so the convergence of $u^K(n)$ is not too fast.

Lemma 5.2.4 *The sequences $(u^K(n))_{n\in\mathbb{N}}$, where $K \in \mathbb{N}$, have the following properties.*

(a) *For any $\varepsilon > 0$, if $K \in \mathbb{N}$ is sufficiently large then*

$$\prod_{n=1}^{\infty} u^K(2^{n-1}) \geq 1 - \frac{\varepsilon}{2}.$$

(b) *If $A \subset \mathbb{N}$ satisfies*

$$\omega(A) := \limsup_{N\to\infty} \frac{1}{N} \cdot |A \cap \{1, \ldots, N\}| > 0$$

then for any $K \in \mathbb{N}$

$$\prod_{n\in A} u^K(n) = 0.$$

Proof.
(a) Let $\varepsilon > 0$. For any sequence $(w^n)_{n\in\mathbb{N}}$ in $[0,1]$ we have

$$\prod_{n=1}^{\infty} w^n = 1 - \left[(1-w^1) + w^1(1-w^2) + w^1w^2(1-w^3) + \cdots\right],$$

thus

$$\begin{aligned}
\prod_{n=1}^{\infty} u^K(2^{n-1}) &= \prod_{n=1}^{\infty} \sqrt[K]{\frac{2^{n-1}}{2^{n-1}+1}} \\
&= \sqrt[K]{\prod_{n=1}^{\infty} \frac{2^{n-1}}{2^{n-1}+1}} \\
&= \sqrt[K]{1 - \left(\frac{1}{2} + \frac{1}{2}\frac{1}{3} + \frac{1}{2}\frac{2}{3}\frac{1}{5} + \frac{1}{2}\frac{2}{3}\frac{4}{5}\frac{1}{9} + \cdots\right)}.
\end{aligned}$$

Let

$$d := 1 - \left(\frac{1}{2} + \frac{1}{2}\frac{1}{3} + \frac{1}{2}\frac{2}{3}\frac{1}{5} + \frac{1}{2}\frac{2}{3}\frac{4}{5}\frac{1}{9} + \cdots\right).$$

Notice that

$$d > 1 - \left(\frac{1}{2} + \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{4} + \frac{1}{2}\frac{1}{8} + \cdots\right) = 0.$$

Since d is positive, there exists a $\bar{K} \in \mathbb{N}$ such that for all $K \geq \bar{K}$

$$\prod_{n=1}^{\infty} u^K(2^{n-1}) = \sqrt[K]{d} \geq 1 - \frac{\varepsilon}{2},$$

so the proof of the first part is complete.

(b) By the definition of $\omega(A)$, there exists an increasing sequence $(n_k)_{k\in\mathbb{N}}$ in A such that

$$\frac{1}{n_k} \cdot |A \cap \{1, \ldots, n_k\}| \geq \frac{1}{2}\omega(A) \qquad \forall k \in \mathbb{N}. \tag{5.2}$$

As $\omega(A) > 0$, by taking a subsequence we may assume without loss of generality that

$$\frac{1}{8} n_{k+1} \cdot \omega(A) \geq n_k \qquad \forall k \in \mathbb{N}. \tag{5.3}$$

Then (5.2) and (5.3) imply

$$\begin{aligned} |A \cap \{n_k + 1, \ldots, n_{k+1}\}| &\geq |A \cap \{1, \ldots, n_{k+1}\}| - n_k \\ &\geq \frac{1}{2} n_{k+1} \cdot \omega(A) - n_k \\ &\geq \frac{1}{4} n_{k+1} \cdot \omega(A). \end{aligned}$$

Since the left hand side is a natural number, we obtain

$$|A \cap \{n_k + 1, \ldots, n_{k+1}\}| \geq \left\lceil \frac{1}{4} n_{k+1} \cdot \omega(A) \right\rceil,$$

where $\lceil r \rceil$ denotes $\min\{n \in \mathbb{N} | r \leq n\}$.
Let $K \in \mathbb{N}$ be arbitrary.. Using that $u^K(n)$ is increasing in n and applying (5.3) yields

$$\begin{aligned} \prod_{n \in A \cap \{n_k+1, \ldots, n_{k+1}\}} u^K(n) &\leq \prod_{n=n_{k+1}-\left\lceil \frac{1}{4} n_{k+1} \cdot \omega(A) \right\rceil + 1}^{n_{k+1}} u^K(n) \\ &= \sqrt[K]{\frac{n_{k+1} - \left\lceil \frac{1}{4} n_{k+1} \cdot \omega(A) \right\rceil + 1}{n_{k+1} + 1}} \\ &\leq \sqrt[K]{\frac{n_{k+1} - \frac{1}{4} n_{k+1} \cdot \omega(A) + 1}{n_{k+1}}} \end{aligned}$$

$$= \sqrt[K]{1 - \frac{1}{4}\omega(A) + \frac{1}{n_{k+1}}}$$
$$\leq \sqrt[K]{1 - \frac{1}{8}\omega(A)}.$$

Therefore

$$\begin{aligned}\prod_{n\in A} u^K(n) &= \prod_{n\in A\cap\{1,\dots,n_1\}} u^K(n) \cdot \prod_{k\in\mathbb{N}} \left[\prod_{n\in(A\cap\{n_k+1,\dots,n_{k+1}\})} u^K(n)\right] \\ &\leq \prod_{k\in\mathbb{N}} \sqrt[K]{1 - \frac{1}{8}\omega(A)} \\ &= 0,\end{aligned}$$

so the proof is complete. □

The next lemma says that, for initial states 1 and 2, if player 2 chooses actions L_1 and L_2 whenever the play is in state 1 or in state 2, then the strategy f^K, with a large $K \in \mathbb{N}$, guarantees that the frequency of visits to state 2 rapidly decreases during the play. At the first sight the reason might seem to be absorption in entry (B_2, L_2), but as it will turn out in lemma 5.2.8, absorption in entry (B_2, L_2) does not play an important role here. The very reason is in fact that the lengths of periods when staying in state 1 increase during the play, which is due to the gradually decreasing probabilities for playing B_1 in state 1.

Lemma 5.2.5 *Let $\varepsilon > 0$, $t \in \{1,2\}$, and let $y = (1,1) \in Y$. For a history $h^\infty \in H_t^\infty$, let $m(h^\infty)$ be the number of stages at which the play is in state 2 during h^∞. Let $M(h^\infty) := \{n \in \mathbb{N} | \, n \leq m(h^\infty)\}$. Let $(a^n(h^\infty))_{n\in M(h^\infty)}$ denote the sequence of stages at which state 2 is visited during h^∞. Then for large $K \in \mathbb{N}$*

$$\mathcal{P}_{tf^K y}\left(a^n(\theta^\infty) \geq 2^{n-1} \quad \forall n \in M(\theta^\infty)\right) \geq 1 - \frac{\varepsilon}{2},$$

where θ^∞ denotes the random variable for the infinite history.

Proof. We only show the statement for initial state 2; for initial state 1 a similar proof can be given. So suppose that the initial state is state 2. Then notice that $a^1(h^\infty) = 1$, $m(h^\infty) \geq 1$, and $M(h^\infty) \neq \emptyset$ for all $h^\infty \in H_2^\infty$ (for initial state 1 we would have that if $M(h^\infty) \neq \emptyset$ then $a^1(h^\infty) \geq 2$, which

would only slightly modify the rest of the proof). For all $h^\infty \in H_2^\infty$, whenever $m(h^\infty) < \infty$, we define inductively

$$a^n(h^\infty) := \max\left\{2^{n-1},\, 8a^{n-1}(h^\infty)\right\} \qquad \forall n \geq m(h^\infty)+1 \tag{5.4}$$

In view of (5.4), we have to show that for large $K \in \mathbb{N}$

$$\mathcal{P}_{2f^K y}\left(a^n(\theta^\infty) \geq 2^{n-1} \quad \forall n \in \mathbb{N}\right) \geq 1 - \frac{\varepsilon}{2}. \tag{5.5}$$

Observe that if the play is in state 2 at stage w, then the probability, with respect to (f^K, y), that the play does not return to state 2 before stage $8w$ is at least the probability that the play moves to state 1 and it stays there till stage $8w-1$; so at least

$$u^K(w) \cdot \prod_{n=w+1}^{8w-2} u^K(n) = \prod_{n=w}^{8w-2} u^K(n).$$

Hence, for any $w, k \in \mathbb{N}$, if $\mathcal{P}_{2f^K y}(a^k(\theta^\infty) = w,\ k \in M(\theta^\infty)) > 0$ then

$$\mathcal{P}_{2f^K y}\left(a^{k+1}(\theta^\infty) \geq 8a^k(\theta^\infty) \,|\, a^k(\theta^\infty) = w,\ k \in M(\theta^\infty)\right) \tag{5.6}$$

$$\geq \prod_{n=w}^{8w-2} u^K(n).$$

On the other hand, if $\mathcal{P}_{2f^K y}(a^k(\theta^\infty) = w,\ k \in \mathbb{N} \setminus M(\theta^\infty)) > 0$ then by (5.4) we have

$$\mathcal{P}_{2f^K y}\left(a^{k+1}(\theta^\infty) \geq 8a^k(\theta^\infty) \,|\, a^k(\theta^\infty) = w,\ k \in \mathbb{N} \setminus M(\theta^\infty)\right) = 1. \tag{5.7}$$

Therefore, for all $w \in \mathbb{N}$ and for all $k \in \mathbb{N}$ satisfying $\mathcal{P}_{2f^K y}(a^k(\theta^\infty) = w) > 0$, by (5.6) and (5.7) we have

$$\begin{aligned}
\mathcal{P}_{2f^K y}\left(a^{k+1}(\theta^\infty) \geq 8a^k(\theta^\infty) \,|\, a^k(\theta^\infty) = w\right) &\geq \prod_{n=w}^{8w-2} u^K(n) \qquad (5.8)\\
&= \sqrt[K]{\frac{w}{(8w-2)+1}}\\
&= \sqrt[K]{\frac{w}{8w-1}}\\
&\geq \sqrt[K]{\frac{1}{8}}.
\end{aligned}$$

For $h^\infty \in H_2^\infty$, let $\eta^0(h^\infty) := 0$ and for $n \in \mathbb{N}$ let

$$\eta^n(h^\infty) := \begin{cases} \eta^{n-1}(h^\infty) + 1 & \text{if } a^{n+1}(h^\infty) \geq 8a^n(h^\infty) \\ \eta^{n-1}(h^\infty) - 1 & \text{otherwise.} \end{cases}$$

We now show that for large $K \in \mathbb{N}$

$$\mathcal{P}_{2f^K y}(\eta^n(\theta^\infty) \geq 1 \quad \forall n \in \mathbb{N}) \geq 1 - \frac{\varepsilon}{2}. \tag{5.9}$$

Let $\xi^K := \sqrt[K]{\frac{1}{8}}$. On the set of integers, for any $K \in \mathbb{N}$, we define a birth and death process $\bar{\eta}_K^n$, $n = 0, 1, 2, \ldots$, as follows. Let $\bar{\eta}_K^0 := 0$ and for $n \in \mathbb{N}$ let

$$\bar{\eta}_K^n := \begin{cases} \bar{\eta}_K^{n-1} + 1 & \text{with probability } \xi^K \\ \bar{\eta}_K^{n-1} - 1 & \text{with probability } 1 - \xi^K. \end{cases}$$

Since ξ^K converges to 1 as K tends to infinity, for the birth and death process $\bar{\eta}_K^n$, $n = 0, 1, 2, \ldots$, we clearly have that for large $K \in \mathbb{N}$

$$\mathcal{P}(\bar{\eta}_K^n \geq 1 \quad \forall n \in \mathbb{N}) \geq 1 - \frac{\varepsilon}{2}.$$

Hence by the definitions of η^n and $\bar{\eta}_K^n$ for $n = 0, 1, 2, \ldots$, and by (5.8) we have for large $K \in \mathbb{N}$ that

$$\mathcal{P}_{2f^K y}(\eta^n(\theta^\infty) \geq 1 \ \forall n \in \mathbb{N}) \ \geq \ \mathcal{P}(\bar{\eta}_K^n \geq 1 \ \forall n \in \mathbb{N}) \ \geq 1 - \frac{\varepsilon}{2},$$

which completes the proof of (5.9).
For $h^\infty \in H_2^\infty$, let $\nu^0(h^\infty) := 0$ and let $\nu^n(h^\infty)$ denote the number of jumps with $+1$ in the sequence $\eta^0(h^\infty)$, $\eta^1(h^\infty), \ldots, \eta^n(h^\infty)$. Since for all $n \in \mathbb{N}$

$$\eta^n(h^\infty) = (+1) \cdot \nu^n(h^\infty) + (-1) \cdot (n - \nu^n(h^\infty)) = 2\nu^n(h^\infty) - n,$$

for large $K \in \mathbb{N}$, inequality (5.9) implies

$$\mathcal{P}_{2f^K y}\left(\nu^n(\theta^\infty) \geq \frac{n+1}{2} \quad \forall n \in \mathbb{N}\right) \geq 1 - \frac{\varepsilon}{2}. \tag{5.10}$$

Recall that $a^1(h^\infty) = 1$ for all $h^\infty \in H_2^\infty$ and notice that if $\nu^n(h^\infty) \geq \frac{n+1}{2}$ for some $n \in \mathbb{N}$, $h^\infty \in H_2^\infty$, then

$$a^n(h^\infty) \geq 8^{\nu^n(h^\infty)-1} \geq 8^{\frac{n-1}{2}} = 2^{\frac{3}{2}(n-1)} \geq 2^{n-1},$$

hence (5.10) implies (5.5), which completes the proof. □

Let $t \in S$, $\pi \in \Pi$, $\sigma \in \Sigma$. Recall the construction and the properties of the probability measure space $(H_t^\infty, \mathcal{S}(\mathcal{M}_t^\infty), \mathcal{P}_{t\pi\sigma})$ in section 2.3.
In the remainder of this chapter, if h^∞ is an infinite history, then h^n denotes the history h^∞ up to stage n. Let

$$H_t^\infty(\pi,\sigma) := \{h^\infty \in H_t^\infty \,|\, \mathcal{P}_{t\pi\sigma}(h^n) > 0 \text{ for all } n \in \mathbb{N}\}.$$

By the definitions, we clearly have $H_t^\infty(\pi,\sigma) \in \mathcal{S}(\mathcal{M}_t^\infty)$ and

$$\mathcal{P}_{t\pi\sigma}(\theta^\infty \in H_t^\infty(\pi,\sigma)) = 1,$$

where θ^∞ denotes the random variable for the infinite history.

If $U_t^\infty \subset H_t^\infty$ for some $t \in S$ then let

$$U_t^n := \{h^n \in H_t^n \,|\, h^n = \bar{h}^n \text{ for some } \bar{h}^\infty \in U_t^\infty\}.$$

By the definitions we obviously have that

$$U_t^\infty \subset \bigcap_{n\in\mathbb{N}} U_t^n = \{h^\infty \in H_t^\infty \,|\, h^n \in U_t^n \quad \text{for all } n \in \mathbb{N}\}.$$

Note that equality above does not need to hold, as illustrated by the following short example. Consider a stochastic game in which there is only one state with two actions T, B for player 1 and one action for player 2. In this game, the set of infinite histories can be seen as the set $H^\infty = \times_{n\in\mathbb{N}} \{T, B\}$. Let U^∞ be the set of all infinite histories which contain action B only finitely many times. Then $U^n = H^n$ for all $n \in \mathbb{N}$, hence $\cap_{n\in\mathbb{N}} U^n = H^\infty \neq U^\infty$ indeed.

Let $U_t^\infty \subset H_t^\infty(\pi,\sigma)$ satisfy the following two properties:

$$U_t^\infty \in \mathcal{S}(\mathcal{M}_t^\infty) \setminus \{\emptyset\} \quad \text{and} \quad U_t^\infty = \bigcap_{n\in\mathbb{N}} U_t^n.$$

For $n \in \mathbb{N}\cup\{0\}$, let $\mathcal{M}_t^n|U_t^\infty$ be the set consisting of all the subsets of U_t^n. Then $(U_t^n, \mathcal{M}_t^n|U_t^\infty)$ are measurable spaces. On these spaces, we define probability measures $\mathcal{P}^n_{t\pi\sigma|U_t^\infty}$ as follows. For $n = 0$ let $\mathcal{P}^0_{t\pi\sigma|U_t^\infty} := \mathcal{P}^0_{t\pi\sigma}$. We proceed inductively. Suppose that $\mathcal{P}^{n-1}_{t\pi\sigma|U_t^\infty}$ is already defined on $(U_t^{n-1}, \mathcal{M}_t^{n-1}|U_t^\infty)$. Then let

$$\mathcal{P}^n_{t\pi\sigma|U_t^\infty}(h^n) := \mathcal{P}^{n-1}_{t\pi\sigma|U_t^\infty}(h^{n-1}) \cdot \mathcal{P}_{t\pi\sigma}(\theta^n = h^n | \theta^n \in U_t^n,\ \theta^{n-1} = h^{n-1})$$

for all $h^n \in U_t^n$, and let

$$\mathcal{P}^n_{t\pi\sigma|U_t^\infty}(V^n) = \sum_{h^n \in V^n} \mathcal{P}^n_{t\pi\sigma|U_t^\infty}(h^n) \qquad \forall V^n \in \mathcal{M}_t^n|U_t^\infty.$$

So we have defined a probability measure $\mathcal{P}^n_{t\pi\sigma|U_t^\infty}$ on $(U_t^n, \mathcal{S}(\mathcal{M}_t^n|U_t^\infty))$. Similarly as in section 2.3, the probability measure spaces

$$(U_t^n, \mathcal{S}(\mathcal{M}_t^n|U_t^\infty), \mathcal{P}^n_{t\pi\sigma|U_t^\infty}), \quad \forall n \in \mathbb{N} \cup \{0\},$$

can be extended to a probability measure space

$$(U_t^\infty, \mathcal{S}(\mathcal{M}_t^\infty|U_t^\infty), \mathcal{P}_{t\pi\sigma|U_t^\infty})$$

in a consistent manner. Note that for this extension the condition $U_t^\infty = \cap_{n\in\mathbb{N}} U_t^n$ is crucial (cf. Dudley [1989], theorem 3.1.1 and theorem 3.1.10). The probability measure $\mathcal{P}_{t\pi\sigma|U_t^\infty}$ corresponds to a stochastic process in which it is assumed that the play does not leave the set U_t^n at any stage n. Note that this probability measure should not be confused with $\mathcal{P}_{t\pi\sigma}(\cdot|\theta^\infty \in U_t^\infty)$. For an illustration, consider the stochastic game, as above, in which there is only one state with two actions T, B for player 1 and one action for player 2. Recall that the set of infinite histories is the set $H^\infty = \times_{n\in\mathbb{N}} \{T, B\}$. Let U^∞ be the set consisting of all infinite histories which start with action T and the infinite history h_B^∞ containing only action B. Take the stationary strategy $x = (1/2, 1/2)$ for player 1, and let y denote player 2's only strategy. Then $\mathcal{P}_{1xy|U^\infty}(h_B^\infty) = 1/2$ even though $\mathcal{P}_{1xy}(h_B^\infty|\theta^\infty \in U^\infty) = 0$.
We remark that

$$\mathcal{S}(\mathcal{M}_t^\infty|U_t^\infty) = \{V_t^\infty \cap \mathcal{U}_t^\infty |\, V_t^\infty \in \mathcal{S}(\mathcal{M}_t^\infty)\}.$$

Next, we present a technical lemma.

Lemma 5.2.6 *Let $t \in S$, $\pi \in \Pi$, $\sigma \in \Sigma$. Let θ^∞ denote the random variable for the infinite history. Let $U_t^\infty \subset H_t^\infty(\pi, \sigma)$ satisfy the following two properties:*

$$U_t^\infty \in \mathcal{S}(\mathcal{M}_t^\infty) \setminus \{\emptyset\} \quad \text{and} \quad U_t^\infty = \bigcap_{n\in\mathbb{N}} U_t^n.$$

For all $h^\infty \in U_t^\infty$ let

$$Z_{t\pi\sigma|U_t^\infty}(h^\infty) := \prod_{k=0}^{\infty} \mathcal{P}_{t\pi\sigma}\left(\theta^{k+1} \in U_t^{k+1} |\, \theta^k = h^k\right).$$

Let the probability measure space $(U_t^\infty, \mathcal{S}(\mathcal{M}_t^\infty | U_t^\infty), \mathcal{P}_{t\pi\sigma|U_t^\infty})$ *be defined as above. Then*

$$\mathcal{P}_{t\pi\sigma}\left(\theta^\infty \in U_t^\infty\right) = \int_{U_t^\infty} Z_{t\pi\sigma|U_t^\infty}\left(h^\infty\right)\, d\mathcal{P}_{t\pi\sigma|U_t^\infty}(h^\infty).$$

Proof. First, for all $n \in \mathbb{N}$ and $h^n \in U_t^n$, let

$$Z^n_{t\pi\sigma|U_t^\infty}\left(h^n\right) := \prod_{k=0}^{n-1} \mathcal{P}_{t\pi\sigma}\left(\theta^{k+1} \in U_t^{k+1} | \, \theta^k = h^k\right).$$

For $h^\infty \in U_t^\infty$ let

$$Z^n_{t\pi\sigma|U_t^\infty}\left(h^\infty\right) := Z^n_{t\pi\sigma|U_t^\infty}\left(h^n\right),$$

hence

$$Z_{t\pi\sigma|U_t^\infty}\left(h^\infty\right) = \lim_{n\to\infty} Z^n_{t\pi\sigma|U_t^\infty}\left(h^\infty\right).$$

For any $h^{k+1} \in U_t^{k+1}$ we have

$$\mathcal{P}_{t\pi\sigma}\left(\theta^{k+1} = h^{k+1} | \, \theta^k = h^k\right) =$$

$$= \mathcal{P}_{t\pi\sigma}\left(\theta^{k+1} \in U_t^{k+1} | \, \theta^k = h^k\right) \cdot \mathcal{P}_{t\pi\sigma|U_t^\infty}(\theta^{k+1} = h^{k+1} | \theta^k = h^k),$$

hence for any $h^n \in U_t^n$ we obtain

$$\mathcal{P}_{t\pi\sigma}\left(\theta^n = h^n\right) =$$

$$= \mathcal{P}_{t\pi\sigma}\left(\theta^1 = h^1 | \, \theta^0 = h^0\right) \cdots \mathcal{P}_{t\pi\sigma}\left(\theta^n = h^n | \, \theta^{n-1} = h^{n-1}\right)$$

$$= Z^n_{t\pi\sigma|U_t^\infty}\left(h^n\right) \cdot \mathcal{P}_{t\pi\sigma|U_t^\infty}(h^n).$$

By the assumptions $U_t^\infty = \cap_{n\in\mathbb{N}} U_t^n$. So using that $U_t^n \supset U_t^{n+1}$ for all $n \in \mathbb{N}$, we must have

$$\mathcal{P}_{t\pi\sigma}\left(\theta^\infty \in U_t^\infty\right) = \lim_{n\to\infty} \mathcal{P}_{t\pi\sigma}\left(\theta^n \in U_t^n\right).$$

Therefore

$$\begin{aligned}
\mathcal{P}_{t\pi\sigma}\left(\theta^\infty \in U_t^\infty\right) &= \lim_{n\to\infty} \mathcal{P}_{t\pi\sigma}\left(\theta^n \in U_t^n\right) \\
&= \lim_{n\to\infty} \sum_{h^n \in U_t^n} \mathcal{P}_{t\pi\sigma}\left(\theta^n = h^n\right) \\
&= \lim_{n\to\infty} \sum_{h^n \in U_t^n} Z^n_{t\pi\sigma|U_t^\infty}\left(h^n\right) \cdot \mathcal{P}_{t\pi\sigma|U_t^\infty}(h^n) \\
&= \lim_{n\to\infty} \int_{U_t^n} Z^n_{t\pi\sigma|U_t^\infty}\left(h^n\right) d\mathcal{P}_{t\pi\sigma|U_t^\infty}(h^n) \\
&= \lim_{n\to\infty} \int_{U_t^\infty} Z^n_{t\pi\sigma|U_t^\infty}\left(h^\infty\right) d\mathcal{P}_{t\pi\sigma|U_t^\infty}(h^\infty) \\
&= \int_{U_t^\infty} Z_{t\pi\sigma|U_t^\infty}\left(h^\infty\right) d\mathcal{P}_{t\pi\sigma|U_t^\infty}(h^\infty),
\end{aligned}$$

where the last equality follows from the monotone convergence theorem (cf. Dudley [1989], section 4.3). Now the proof is complete. □

Note that, in the game Γ, the history up to any stage $n \in \mathbb{N}$ already determines the state at stage $n+1$, because for each action pair, the transition occurs to a certain state with probability 1. Therefore, in the game Γ, the mixed actions prescribed by the strategies at stage $n+1$ are already determined by the history up to stage n.
The following result, which will follow from the previous lemma, intuitively states that the set of infinite histories in which absorption should occur with probability 1 but no absorption occurs has probability zero.

Lemma 5.2.7 *Let $t \in \{1,2\}$, $K \in \mathbb{N}$, $\sigma \in \Sigma^p$. Let*

$$\tilde{H}_t^\infty := \{h^\infty \in H_t^\infty(f^K, \sigma) |\ \textit{no absorption occurs in } h^\infty\}$$

$$\bar{H}_t^\infty := \{h^\infty \in \tilde{H}_t^\infty |\ \prod_{n \in C(h^\infty)} u^K(n) = 0\},$$

where $C(h^\infty)$ is the set of stages n when, according to the pure strategy σ, player 2 plays actions R_1, R_2, or L_2 after history h^{n-1}. Let θ^∞ denote the random variable for the infinite history. Then

$$\mathcal{P}_{tf^K\sigma}\left(\theta^\infty \in \bar{H}_t^\infty\right) = 0.$$

Proof. Let $N \in \mathbb{N}$ and $\delta > 0$. Let

$$V_t^\infty(N,\delta) := \{h^\infty \in \tilde{H}_t^\infty | \prod_{n \in C(h^\infty) \cap \{1,\dots,N\}} u^K(n) \leq \delta\}.$$

It follows from the definitions that $V_t^\infty(N,\delta) \in \mathcal{S}(\mathcal{M}_t^\infty) \setminus \{\emptyset\}$ and $V_t^\infty(N,\delta) = \bigcap_{n\in\mathbb{N}} V_t^n(N,\delta)$. Let $Z_{tf^K\sigma|V_t^\infty(N,\delta)}$ be defined as in lemma 5.2.6. Then we have for all $h^\infty \in V_t^\infty(N,\delta)$

$$\begin{aligned} Z_{tf^K\sigma|V_t^\infty(N,\delta)}(h^\infty) &= \prod_{k=0}^{\infty} \mathcal{P}_{tf^K\sigma}\left(\theta^{k+1} \in V_t^{k+1}(N,\delta) | \, \theta^k = h^k\right) \\ &\leq \prod_{k=0}^{\infty} \mathcal{P}_{tf^K\sigma}\left(\theta^{k+1} \in \tilde{H}_t^{k+1} | \, \theta^k = h^k\right) \\ &= \prod_{n \in C(h^\infty)} u^K(n) \\ &\leq \delta. \end{aligned}$$

Hence lemma 5.2.6 implies

$$\mathcal{P}_{tf^K\sigma}\left(\theta^\infty \in V_t^\infty(N,\delta)\right) \leq \delta. \tag{5.11}$$

Let

$$V_t^\infty(\delta) := \{h^\infty \in \tilde{H}_t^\infty | \prod_{n \in C(h^\infty)} u^K(n) < \delta\}.$$

By using

$$V_t^\infty(N,\delta) \subset V_t^\infty(N+1,\delta) \qquad \forall N \in \mathbb{N}$$

$$V_t^\infty(\delta) \subset \bigcup_{N\in\mathbb{N}} V_t^\infty(N,\delta),$$

inequality (5.11) yields

$$\mathcal{P}_{tf^K\sigma}\left(\theta^\infty \in V_t^\infty(\delta)\right) \leq \delta.$$

As $\delta > 0$ was arbitrary, the inclusion $\bar{H}_t^\infty \subset V_t^\infty(\delta)$ completes the proof. □

It turns out that the strategy f^K, with a large $K \in \mathbb{N}$, keeps the probability of absorption in entry (B_2, L_2) small. In fact, the absorption probability in entry (B_2, L_2) is maximal when player 2 always chooses actions L_1 and L_2

whenever the play is in state 1 or in state 2. But even then, in view of lemma 5.2.5, the play does not visit state 2 "frequently enough", so using that f^K assigns less and less probabilities to action B_2, the probability of absorption in entry (B_2, L_2) turns out to be small indeed.

Lemma 5.2.8 *Let $\varepsilon > 0$. If $K \in \mathbb{N}$ is sufficiently large then, for initial states $t = 1, 2$ in the game Γ, the probability of absorption in entry (B_2, L_2) is at most ε with respect to (f^K, σ), for any $\sigma \in \Sigma$.*

Proof. It is easy to see that the stationary strategy y which chooses actions L_1 and L_2 with probability 1 maximizes the probability of absorption in entry (B_2, L_2) against f^K, with any $K \in \mathbb{N}$. Therefore it suffices to show the statement for y.
We only show the statement for initial state 2. Then for initial state 1 the statement is immediate, since from the stage on when the play moves to state 2, the strategy f^K assigns even less probabilities to actions B_1 and B_2 than when starting from stage 1. So assume the initial state is state 2. Let

$$\begin{aligned} \tilde{H}_2^\infty &: = \{h^\infty \in H_2^\infty(f^K, y) | \text{ no absorption occurs in } h^\infty\} \\ \hat{H}_2^\infty &: = \{h^\infty \in H_2^\infty(f^K, y) | \, a^n(h^\infty) \geq 2^{n-1} \quad \forall n \in M(h^\infty)\}, \end{aligned}$$

where $a^n(h^\infty)$ and $M(h^\infty)$ are defined as in lemma 5.2.5. It is clear by the definitions that $\hat{H}_2^\infty \in \mathcal{S}(\mathcal{M}_2^\infty) \setminus \{\emptyset\}$ and $\hat{H}_2^\infty = \cap_{n \in \mathbb{N}} \hat{H}_2^n$. We define a probability measure on $(H_2^\infty, \mathcal{S}(\mathcal{M}_2^\infty))$ as follows:

$$\mathcal{P}^*_{2f^K y}(M) := \mathcal{P}_{2f^K y | \hat{H}_2^\infty}(M \cap \hat{H}_2^\infty) \qquad \forall M \in \mathcal{S}(\mathcal{M}_2^\infty).$$

For any $\delta > 0$, if $K \in \mathbb{N}$ is large then by lemma 5.2.5 we have

$$\mathcal{P}_{2f^K y}\left(\theta^\infty \in \hat{H}_2^\infty\right) \geq 1 - \delta.$$

One can show that these inequalities imply that for any $\delta > 0$, if $K \in \mathbb{N}$ is large then

$$|\mathcal{P}_{2f^K y}(M) - \mathcal{P}^*_{2f^K y}(M)| \leq \delta \qquad \forall M \in \mathcal{S}(\mathcal{M}_2^\infty). \tag{5.12}$$

Let $U_2^\infty := \tilde{H}_2^\infty \cap \hat{H}_2^\infty$. It is clear that

$$U_2^\infty \in \mathcal{S}(\mathcal{M}_2^\infty) \setminus \{\emptyset\}, \quad U_2^\infty = \cap_{n \in \mathbb{N}} U_2^n, \quad U_2^n := \tilde{H}_2^n \cap \hat{H}_2^n \quad \forall n \in \mathbb{N}.$$

We will now use lemma 5.2.6 for the above defined probability measure $\mathcal{P}^*_{2f^K y}$ instead of $\mathcal{P}_{2f^K y}$. Now for $h^\infty \in U_2^\infty$ let $Z^*_{2f^K y | U_2^\infty}(h^\infty)$ be defined for $\mathcal{P}^*_{2f^K y}$ as in lemma 5.2.6.

Then, using lemma 5.2.4-(a), if $K \in \mathbb{N}$ is sufficiently large then for any $h^\infty \in U_2^\infty$

$$\begin{aligned} Z^*_{2f^K y|U_2^\infty}(h^\infty) &= \prod_{k=0}^{\infty} \mathcal{P}^*_{2f^K y}\left(\theta^{k+1} \in U_2^{k+1} | \, \theta^k = h^k\right) \\ &= \prod_{k=0}^{\infty} \mathcal{P}^*_{2f^K y}\left(\theta^{k+1} \in \tilde{H}_2^{k+1} | \, \theta^k = h^k\right) \\ &= \prod_{n \in M(h^\infty)} u^K(a^n(h^\infty)) \\ &\geq \prod_{n \in M(h^\infty)} u^K(2^{n-1}) \\ &\geq \prod_{n=1}^{\infty} u^K(2^{n-1}) \\ &\geq 1 - \frac{\varepsilon}{2}. \end{aligned}$$

Hence, by applying (5.12) and lemma 5.2.6, for large $K \in \mathbb{N}$ we get

$$\begin{aligned} \mathcal{P}_{2f^K y}\left(\theta^\infty \in \tilde{H}_2^\infty\right) &\geq \mathcal{P}^*_{2f^K y}(\theta^\infty \in U_2^\infty) - \frac{\varepsilon}{2} \\ &= \int_{U_2^\infty} Z^*_{2f^K y|U_2^\infty}(h^\infty) \, d\mathcal{P}^*_{2f^K y}(h^\infty) - \frac{\varepsilon}{2} \\ &\geq 1 - \frac{\varepsilon}{2} - \frac{\varepsilon}{2} \\ &= 1 - \varepsilon, \end{aligned}$$

which means that if $K \in \mathbb{N}$ is large then, with respect to (f^K, y) and initial state 2, the probability of absorption in entry (B_2, L_2) is at most ε. □

Now we show that, when player 1 uses f^K with any $K \in \mathbb{N}$ for initial states 1 or 2, if player 2 chooses actions R_1 and R_2 "too frequently" then absorption occurs with probability 1. (Recall again that, in the game Γ, the history up to any stage $n \in \mathbb{N}$ already determines the state for stage $n+1$.)

Lemma 5.2.9 *Let $t \in \{1,2\}$, $K \in \mathbb{N}$, $\sigma \in \Sigma^p$. Let*

$$\tilde{H}_t^\infty := \{h^\infty \in H_t^\infty(f^K, \sigma) | \, \textit{no absorption occurs in } h^\infty\}.$$

For $A \subset \mathbb{N}$ let

$$\omega(A) := \limsup_{N \to \infty} \frac{1}{N} \cdot |A \cap \{1, \ldots, N\}|.$$

For a history $h^\infty \in H_t^\infty$, let $A(h^\infty)$ denote the set of stages n when, according to the pure strategy σ, player 2 chooses actions R_1 or R_2 after history h^{n-1}. Then, if

$$\mathcal{P}_{tf^K\sigma}\left(\theta^\infty \in \tilde{H}_t^\infty\right) > 0$$

then

$$\mathcal{P}_{tf^K\sigma}\left(\omega(A(\theta^\infty)) = 0 | \, \theta^\infty \in \tilde{H}_t^\infty\right) = 1,$$

where θ^∞ denotes the random variable for the infinite history.

Proof. Suppose that $\omega(A(h^\infty)) > 0$ for some history $h^\infty \in H_t^\infty$. Then, clearly, no absorption occurs in h^∞, thus $h^\infty \in \tilde{H}_t^\infty$. Let

$$\bar{H}_t^\infty := \{h^\infty \in \tilde{H}_t^\infty | \prod_{n \in A(h^\infty) \cup B(h^\infty)} u^K(n) = 0\},$$

where $B(h^\infty)$ is the set of stages n when, according to the pure strategy σ, player 2 plays action L_2 after history h^{n-1}. By lemma 5.2.4-(b) we have

$$\prod_{n \in A(h^\infty)} u^K(n) = 0,$$

therefore

$$\{h^\infty \in H_t^\infty(f^K, \sigma) | \, \omega(A(h^\infty)) > 0\} \subset \bar{H}_t^\infty.$$

Now lemma 5.2.7 yields

$$\mathcal{P}_{tf^K\sigma}\left(\omega(A(\theta^\infty)) > 0\right) \leq \mathcal{P}_{tf^K\sigma}\left(\theta^\infty \in \bar{H}_t^\infty\right) = 0,$$

which implies the statement. □

The next result tells us that, when player 1 uses f^K with any $K \in \mathbb{N}$ for initial states 1 and 2, then, given that no absorption occurs (and this has a positive probability), the average of the payoffs along the infinite history equals 1 almost surely.

Lemma 5.2.10 *Let $t \in \{1, 2\}$, $K \in \mathbb{N}$, $\sigma \in \Sigma^p$. Let*

$$\tilde{H}_t^\infty := \{h^\infty \in H_t^\infty(f^K, \sigma) | \, \textit{no absorption occurs in } h^\infty\}.$$

Then, if

$$\mathcal{P}_{tf^K\sigma}\left(\theta^\infty \in \tilde{H}_t^\infty\right) > 0$$

then

$$\mathcal{P}_{tf^K\sigma}\left(\liminf_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} R_n = 1 \,|\, \theta^\infty \in \tilde{H}_t^\infty\right) = 1,$$

where θ^∞ denotes the random variable for the infinite history and R_n the random variable for the payoff at stage n.

Proof. Let $\omega(A)$ for $A \subset \mathbb{N}$ and $A(h^\infty)$ be defined as in lemma 5.2.9. Let $R_n(h^\infty)$ be the payoff at stage n according to the history h^∞. Clearly, we have $R_n = R_n(\theta^\infty)$.
Now for any $h^\infty \in \tilde{H}_t^\infty$

$$\begin{aligned}
\liminf_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} R_n(h^\infty) &= \lim_{m\to\infty} \inf_{N\geq m} \frac{\sum_{n=1}^{N} R_n(h^\infty)}{N} \\
&= \lim_{m\to\infty} \inf_{N\geq m} \frac{|\{n \in \{1,\ldots,N\} \,|\, R_n(h^\infty) = 1\}|}{N} \\
&= \lim_{m\to\infty} \inf_{N\geq m} \frac{N - |\{n \in \{1,\ldots,N\} \,|\, R_n(h^\infty) = 0\}|}{N} \\
&= \lim_{m\to\infty} \inf_{N\geq m} \frac{N - |A(h^\infty) \cap \{1,\ldots,N\}|}{N} \\
&= 1 + \lim_{m\to\infty} \inf_{N\geq m} \frac{-|A(h^\infty) \cap \{1,\ldots,N\}|}{N} \\
&= 1 - \lim_{m\to\infty} \sup_{N\geq m} \frac{|A(h^\infty) \cap \{1,\ldots,N\}|}{N} \\
&= 1 - \omega(A(h^\infty)),
\end{aligned}$$

hence lemma 5.2.9 implies the result. □

Now we are ready to prove that $\mathcal{B}_t = 1$ for initial states $t = 1, 2$ and also that the Markov strategy f^K is ε-optimal for large $K \in \mathbb{N}$. More specifically, K can be any number that satisfies lemma 5.2.8.

Lemma 5.2.11 *Let $t \in \{1,2\}$. Then*

$$\mathcal{B}_t = v_t = 1.$$

Furthermore, for any $\varepsilon > 0$, if $K \in \mathbb{N}$ is sufficiently large then

$$\underline{v}_t(f^K) \geq 1 - \varepsilon.$$

Proof. Let $t \in \{1,2\}$ and let $\varepsilon > 0$. We only need to show that $\underline{v}_t\left(f^K\right) \geq 1-\varepsilon$ for large $K \in \mathbb{N}$, because then $\mathcal{B}_t = v_t = 1$ follows from (5.1) and from the fact that the largest payoff in the game is 1. Let θ^∞ denote the random variable for the infinite history and R_n the random variable for the payoff at stage n. By lemma 5.2.10 we have with respect to (f^K, σ), for any $K \in \mathbb{N}$ and for any $\sigma \in \Sigma^p$, and initial state t with probability 1 that

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} R_n = \begin{cases} 0 & \text{if absorption occurs in entry } (B_2, L_2) \text{ in } \theta^\infty \\ 1 & \text{otherwise.} \end{cases}$$

Take $K \in \mathbb{N}$ as in lemma 5.2.8. Then the probability of absorption in entry (B_2, L_2) is at most ε with respect to (f^K, σ) and initial state t, for any $\sigma \in \Sigma^p$, hence

$$\mathcal{E}_{tf^K\sigma}\left(\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} R_n\right) \geq 1 - \varepsilon \qquad \forall \sigma \in \Sigma^p. \tag{5.13}$$

By applying Fatou's lemma (cf. Fatou [1906]) we obtain for all $\sigma \in \Sigma^p$ that

$$\begin{aligned} \gamma_t(f^K, \sigma) &= \liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathcal{E}_{tf^K\sigma}(R_n) \\ &\geq \mathcal{E}_{tf^K\sigma}\left(\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} R_n\right) \\ &\geq 1 - \varepsilon. \end{aligned}$$

In view of theorem 2.8.2-(a), it suffices to consider pure replies from player 2, thus

$$\underline{v}_t(f^K) \geq 1 - \varepsilon,$$

so the proof is complete. □

5.3 Sufficient conditions for $\mathcal{A} = \mathcal{B}$

Example 5.2.1 in the previous section demonstrated that $\mathcal{B}$ may be strictly larger than $\mathcal{A}$ for some initial states. However, this cannot hold for all initial states, as stated in the next theorem.

Theorem 5.3.1 *In every zero-sum stochastic game* $\mathcal{A}_s = \mathcal{B}_s\,(= v_s)$ *for all states* $s \in S^{\min} := \{t \in S | v_t = \min_{w \in S} v_w\}$.

Proof. Let $s \in S^{\min}$. Then by the results of Thuijsman & Vrieze [1993], for any $\varepsilon > 0$, player 1 has a stationary ε-optimal strategy x^ε for initial state s. Hence $\mathcal{A}_s = v_s$, thus (5.1) yields $\mathcal{A}_s = \mathcal{B}_s\,(= v_s)$. □

We know by now that $\mathcal{A}$ may be strictly smaller than $\mathcal{B}$ and also that $\mathcal{A}$ equals $\mathcal{B}$ for at least one initial state in every zero-sum game. Now the question is what conditions would guarantee that $\mathcal{A}$ equals $\mathcal{B}$ for all initial states. We will present several sufficient conditions, however, first we would like to recall some of the most important classes of zero-sum games in which $\mathcal{A} = \mathcal{B}$ is already known.
Clearly, we have $\mathcal{A} = \mathcal{B}(= v)$ for any class of games where player 1 has stationary ε-optimal strategies, for all $\varepsilon > 0$. The existence of stationary ε-optimal strategies is known in several special classes of stochastic games (cf. theorem 2.11.2). Moreover, the condition that the value is constant ($v_s = v_t$ for all $s, t \in S$) is also sufficient for the existence of stationary ε-optimal strategies (cf. Thuijsman & Vrieze [1993]). (In such games actually both players have Markov optimal strategies as well.)

$\mathcal{A} = \mathcal{B}$ in repeated games with absorbing states

Repeated games with absorbing states are stochastic games where there is only one non-absorbing state. Kohlberg [1974] showed that these games have a value (cf. theorem 2.11.2-(g)). However, to achieve this value history dependent strategies are indispensable. We will show for these games that $\mathcal{A} = \mathcal{B}$. For the specific case of the Big Match we have already shown this equality (cf. lemma 2.9.7-(e)). In fact, we generalize that proof to all repeated games with absorbing states. We will not discuss all the technical details in the proof, we only give a brief sketch. We wish to mention that the same result also follows from Coulomb [1992], who used quite similar techniques in his proof.

Theorem 5.3.2 *In any zero-sum repeated game with absorbing states* $\mathcal{A} = \mathcal{B}$.

Proof. Take a zero-sum repeated game with absorbing states. We may suppose without loss of generality that, in each absorbing state, both players have only one action (otherwise we may replace the state by another absorbing state containing only the value of the original state as payoff). Suppose that the initial state is state 1, the only non-absorbing state. We will often suppress state 1 in the notations.

Any action of player 1 or player 2 in state 1 will also denote the stationary strategy which prescribes this action for each stage.

In view of theorem 2.8.2-(b), against any stationary strategy $x \in X$ there exists a best reply $j^x \in J$, hence we have $\gamma(x, j^x) \leq \mathcal{A}$. This means that, for initial state 1, either (x, j^x) is absorbing (namely it eventually leads to absorption with probability 1, or equivalently, $p_1(1|x, j^x) < 1$) and then the expected absorption payoff is at most $\mathcal{A}_1$, or (x, j^x) is non-absorbing and then the expected one-shot payoff $r_1(x, j^x)$ is at most $\mathcal{A}_1$.

Take an arbitrary Markov strategy $f = (x^n)_{n\in\mathbb{N}} \in F$. Let $\varepsilon > 0$. It suffices to show that there exists a Markov strategy $g \in G$ such that $\gamma_1(f, g) \leq \mathcal{A}_1 + \varepsilon$. Take an arbitrary $\delta > 0$.

Step 1. Let $f_1 := f$ and $x_1^n := x^n$ for all $n \in \mathbb{N}$. Let $g_1 = (j^{x_1^n})_{n\in\mathbb{N}}$. Let ξ denote the random variable for the stage when absorption occurs, if no absorption occurs at all then let $\xi = 0$. For $N \in \mathbb{N} \cup \{0\}$ let

$$p_1^N := \mathcal{P}_{f_1 g_1}(\xi > N),$$

so p_1^N is the probability of absorption after stage N with respect to (f_1, g_1). Let $p_1 := p_1^0$. Clearly, p_1 is the probability of absorption with respect to (f_1, g_1). We have

$$\begin{aligned} p_1 &= \mathcal{P}_{f_1 g_1}(1 \leq \xi) \\ &= \lim_{N\to\infty} \mathcal{P}_{f_1 g_1}(1 \leq \xi \leq N) \\ &= \lim_{N\to\infty} [\mathcal{P}_{f_1 g_1}(1 \leq \xi) - \mathcal{P}_{f_1 g_1}(\xi > N)] \\ &= p_1 - \lim_{N\to\infty} p_1^N, \end{aligned}$$

hence p_1^N converges to 0. If there exists a $N \in \mathbb{N}$ such that $p_1^N = 0$ then let $N_1 := N$. Otherwise, choose a stage N_1 such that $p_1^{N_1} \leq p^* \cdot \delta$, where p^* is the smallest positive absorption probability in state 1:

$$p^* := \min\{p_{ij}^* | \, i \in I, \; j \in J, \; p_{ij}^* := 1 - p_1(1|i, j) \text{ and } p_{ij}^* > 0\};$$

(note that there must exist $i \in I$ and $j \in J$ such that $p_1(1|i, j) < 1$, otherwise state 1 would be absorbing).

If $p_1^{N_1} = 0$ then we have $\gamma_1(f, g_1) = \gamma_1(f_1, g_1) \leq \mathcal{A}_1$, because, with respect to (f_1, g_1), the expected absorption payoff is at most $\mathcal{A}_1$ at each stage $n \leq N_1$; the probability of absorption after stage N_1 is zero; and the expected payoff in state 1 is at most $\mathcal{A}_1$ at each stage $n > N_1$.
Assume now that $p_1^{N_1} > 0$. By the definition of N_1, the probability of absorption after stage N_1 for (f_1, g_1) is at most $p^* \cdot \delta$. Now let $I_1^n := \{i \in I | (i, j^{x_1^n}) \text{ is non-absorbing}\}$. Thus the probability that, with respect to (f_1, g_1), player 1 will ever choose an action outside I_1^n at stages $n > N_1$ is at most δ.

Step 2. Let $x_2^n := x_1^n$ for $n \leq N_1$ and let x_2^n be the normalization of x_1^n on I_1^n for $n > N_1$:

$$x_2^n(i) := \frac{x_1^n(i)}{\sum_{i \in I_1^n} x_1^n(i)} \quad \text{for all } i \in I_1^n, \qquad x_2^n(i) := 0 \quad \text{for all } i \in I \setminus I_1^n.$$

Let $f_2 := (x_2^n)_{n \in \mathbb{N}}$. Intuitively, f_2 coincides with f_1 up to stage N_1, and, after stage N_1, the strategy f_2 equals the strategy f_1 on condition that no action outside I_1^n will ever be chosen at stages $n > N_1$. Let $g_2 := (j^{x_2^n})_{n \in \mathbb{N}}$, so by the definitions, g_1 and g_2 are the same for the first N_1 stages. One can show, using the properties of the construction, that, with respect to (f, g_2), the probability of absorption outside I_1^n at stages $n > N_1$ is at most δ. Similarly to step 1, choose an $N_2 > N_1$ such that $p_2^{N_2} = 0$ if possible, otherwise

$$p_2^{N_2} := \mathcal{P}_{f_2 g_2}(\xi > N_2) \leq \delta \cdot p^*.$$

Assume first that $p_2^{N_2} = 0$. Then we have $\gamma_1(f, g_2) \leq \mathcal{A}_1 + \varepsilon$ for small δ, because, with respect to (f, g_2), the expected absorption payoff at each stage in $n \leq N_1$ is at most $\mathcal{A}_1$; the probability of absorption outside I_1^n at stages $n = N_1 + 1, \ldots, N_2$ is at most δ; the expected absorption payoff in I_1^n at each stage in $n = N_1 + 1, \ldots, N_2$ is at most $\mathcal{A}_1$; the probability of absorption after stage N_2 is zero; and the expected payoff in state 1 at each stage in $n > N_2$ is at most $\mathcal{A}_1$.
Assume now that $p_2^{N_2} > 0$. Let $I_2^n := \{i \in I | (i, j^{x_2^n}) \text{ is non-absorbing}\}$, and repeat the above steps, in such a way that $N_{k+1} > N_k$ for all k, until at some step K we have $p_K^{N_K} = 0$. This results in a strategy g_K for player 2. Note that for $p_K^{N_K} = 0$ it is sufficient that $I_K^n = I_{K-1}^n$ holds for all $n > N_K$. Hence we only need at most $K \leq |I|$ steps because, for any stage $n > N_{k+1}$, either I_{k+1}^n becomes smaller than I_k^n, or $I_k^n = I_{k+1}^n$ and then nothing changes at further steps for stage n. Using similar arguments as before, one can show now that $p_K^{N_K} = 0$ implies that $\gamma_1(f, g_K) \leq \mathcal{A}_1 + \varepsilon$ if $\delta > 0$ is small enough. □

$\mathcal{A} = \mathcal{B}$ in games with constant $\mathcal{A}$ or $\mathcal{B}$

In this section we show that $\mathcal{A} = \mathcal{B}$ in games where $\mathcal{A}$ or $\mathcal{B}$ is constant. We need the following lemma.

Lemma 5.3.3 *Let $\varepsilon > 0$. For $\beta \in (0,1)$, let $x_\beta \in X$ be a β-discounted optimal strategy. Then for large $\beta \in (0,1)$*

$$\underline{v}_s(x_\beta) \geq \min_{t \in S} v_t - \varepsilon \qquad \forall s \in S.$$

Proof. By theorem 2.8.2-(b) and by the finiteness of the state space S and the space J of pure stationary strategies for player 2, it suffices to show that for any $s \in S$ and $j \in J$, if $\beta \in (0,1)$ is large, then

$$\gamma_s(x_\beta, j) \geq \min_{t \in S} v_t - \varepsilon.$$

Let $s \in S$ and $j \in J$. Using theorem 2.9.3 we have

$$(1-\beta) \cdot r(x_\beta, j) + \beta \cdot P(x_\beta, j) \cdot v_\beta \geq v_\beta \qquad \forall \beta \in (0,1).$$

By (2.1), multiplying this inequality with $Q(x_\beta, j)$ yields

$$Q(x_\beta, j) \cdot r(x_\beta, j) \geq Q(x_\beta, j) \cdot v_\beta \qquad \forall \beta \in (0,1).$$

Using theorem 2.7.1-(a) and theorem 2.9.5, we have for large $\beta \in (0,1)$ that

$$\begin{aligned} \gamma_s(x_\beta, j) &= \sum_{t \in S} q_s(t|x_\beta, j)\, r_t(x_{\beta t}, j_t) \\ &\geq \sum_{t \in S} q_s(t|x_\beta, j)\, v_{\beta t} \\ &\geq \sum_{t \in S} q_s(t|x_\beta, j)\, (v_t - \varepsilon) \\ &\geq \min_{t \in S} v_t - \varepsilon, \end{aligned}$$

so the proof is complete. □

With the help of the above lemma we show the following result.

Theorem 5.3.4 *In every zero-sum stochastic game*

$$\min_{s \in S} \mathcal{A}_s = \min_{s \in S} \mathcal{B}_s = \min_{s \in S} v_s, \qquad \max_{s \in S} \mathcal{A}_s = \max_{s \in S} \mathcal{B}_s = \max_{s \in S} v_s.$$

Proof. By lemma 5.3.3, for any $\varepsilon > 0$ player 1 has a stationary strategy x^ε satisfying

$$\underline{v}_t(x^\varepsilon) \geq \min_{s \in S} v_s - \varepsilon \qquad \forall t \in S,$$

hence

$$\min_{s \in S} \mathcal{A}_s \geq \min_{s \in S} v_s,$$

which, in view of (5.1), implies the first part of the statement.
By the results of Thuijsman & Vrieze [1993], there is always a state t in $S^{\max} := \{s \in S | \, v_s = \max_{w \in S} v_w\}$ for which player 1 has a stationary optimal strategy x. Hence

$$\max_{s \in S} \mathcal{A}_s \geq \mathcal{A}_t \geq \underline{v}_t(x) = v_t = \max_{s \in S} v_s,$$

thus (5.1) implies the second part of the statement. □

The above theorem yields the following corollary.

Corollary 5.3.5 *In every zero-sum stochastic game where any of $\mathcal{A}$, $\mathcal{B}$ or v is constant, $\mathcal{A} = \mathcal{B}(= v)$ is constant.*

The following theorem provides a more relaxed view on constant values.

Theorem 5.3.6 *In every zero-sum stochastic game where for all $s, t \in S$ either $\mathcal{A}_s = \mathcal{A}_t$ or $\mathcal{B}_s = \mathcal{B}_t$, we have that $\mathcal{A} = \mathcal{B}(= v)$ is constant.*

Proof. Using the inequality $\mathcal{A} \leq \mathcal{B}$ and theorem 5.3.4, it is clear that if state s has the property that $\mathcal{B}_s = \min_{w \in S} \mathcal{B}_w$ then $\mathcal{A}_s = \min_{w \in S} \mathcal{A}_w$. Similarly, if state t has the property that $\mathcal{A}_t = \max_{w \in S} \mathcal{A}_w$ then $\mathcal{B}_t = \max_{w \in S} \mathcal{B}_w$. Now by the condition we have either $\mathcal{A}_s = \mathcal{A}_t$ or $\mathcal{B}_s = \mathcal{B}_t$. Therefore by theorem 5.3.4 either $\mathcal{A}$ or $\mathcal{B}$ is constant, thus corollary 5.3.5 completes the proof. □

An interesting equivalent formulation of theorem 5.3.6 is the following: if $\mathcal{A} \neq \mathcal{B}$ then there must exist two states s and t such that $\mathcal{A}_s \neq \mathcal{A}_t$ and $\mathcal{B}_s \neq \mathcal{B}_t$.

$\mathcal{A} = \mathcal{B}$ in games with optimal or best-Markov strategies

By a best-Markov strategy we mean a Markov strategy f with the property that $\underline{v}(f) \geq \underline{v}(\bar{f})$ for all $\bar{f} \in F$, or equivalently $\underline{v}(f) = \mathcal{B}$. Optimal strategies and best-Markov strategies do not necessarily exist, but if they do then their existence surprisingly implies $\mathcal{A} = \mathcal{B}$, as stated in the next theorem.

Theorem 5.3.7 *In every zero-sum stochastic game, if player 1 has an optimal strategy or a best-Markov strategy then $\mathcal{A} = \mathcal{B}$.*

Proof. Suppose first that player 1 has an optimal strategy. Then by Main Theorem 3, player 1 has stationary ε-optimal strategies for all $\varepsilon > 0$ and Markov optimal strategies as well, hence (5.1) yields the result.
Assume now that player 1 has a best-Markov strategy f, so $\underline{v}(f) \geq \underline{v}(\bar{f})$ for all $\bar{f} \in F$. Since, for any history $h \in H$, the strategy $f[h]$ is also a Markov strategy, we have $\underline{v}(f) \geq \underline{v}(f[h])$ for all $h \in H$. Hence f must be a non-improving strategy (cf. definition 4.1.1-(b)). Then by Main Theorem 4, for all $\varepsilon > 0$, player 1 has stationary strategies guaranteeing $\underline{v}_s(f) - \varepsilon = \mathcal{B}_s - \varepsilon$ for all initial states s, hence $\mathcal{A}_s \geq \mathcal{B}_s$. Now (5.1) completes the proof. □

Note that in example 5.2.1 player 1 has neither optimal strategies nor best-Markov strategies for initial states 1 and 2. We only show it for initial state 2. One can argue as follows. Since $\mathcal{B}_2 = v_2 = 1$ in that game, it suffices to show that player 1 has no strategy guaranteeing reward 1 for initial state 2. Assume by way of contradiction that a strategy π guarantees 1 for initial state 2. As the largest payoff in the game is 1, π has to prescribe action T_2 with probability 1 whenever the play is in state 2 (otherwise the probability of absorption in entry (B_2, L_2) with payoff 0 would be positive if player 2 chooses action L_2). Thus if player 2 always plays action R_2 in state 2, then the reward is 0, which is a contradiction. Therefore, player 1 has neither optimal nor best-Markov strategies for initial state 2 indeed.

5.4 Concluding remarks

By the definition of $\mathcal{A}$, for each $s \in S$ and for any $\delta > 0$, player 1 has a stationary strategy $x^{s\delta} \in X$ such that $\underline{v}_s(x^{s\delta}) \geq \mathcal{A}_s - \delta$. In this finite state model, it can be shown however that for any $\delta > 0$ we can take $x^{s\delta}$ independent of the initial state, so for all $\delta > 0$ there exists an $x^\delta \in X$ such

that $\underline{v}_s(x^\delta) \geq \mathcal{A}_s - \delta$ for all $s \in S$. This means that the following equality for stationary strategies makes sense:

$$\mathcal{A} = \sup_{x \in X} \underline{v}(x).$$

So we could have used this state independent equality as the definition of $\mathcal{A}$ as well. Note that for games with countable state spaces this equivalence of definitions is not valid. Nowak & Raghavan [1991] presented a game with a countable state space, where even though, for each initial state, player 1 has stationary ε-optimal strategies for all $\varepsilon > 0$, he has no stationary strategies that are ε-optimal for all initial states if ε is small.
Finally, we wish to remark that it is not known whether or not $\mathcal{B}$ can be defined state independently.

Chapter 6

Almost stationary ε-equilibria

6.1 Introduction

In the literature of stochastic games, existence of (ε)-equilibria has been frequently established in terms of almost stationary strategy pairs (see for example Vrieze & Thuijsman [1989], Vieille [1993] or, in a more general fashion, Thuijsman & Vrieze [1996]). Intuitively, a pair of strategies (π, σ) is called almost stationary if there exists a pair of stationary strategies (x^*, y^*) with the following property: with regard to (π, σ) and any initial state, the mixed actions corresponding to these stationary strategies x^* and y^* are used by the players during the whole play, with probability almost 1. So a pair of almost stationary strategies behaves as if it was a pair of simple stationary strategies. Formally, the concept of almost stationary ε-equilibria is defined as follows.

Definition 6.1.1 *An ε-equilibrium (π, σ) is an almost stationary ε-equilibrium, $\varepsilon \geq 0$, if there exists a pair of stationary strategies (x^*, y^*) such that*

$$\mathcal{P}_{s\pi\sigma}\left(\pi_{s^n}(\theta^{n-1}) = x^*_{s^n},\ \sigma_{s^n}(\theta^{n-1}) = y^*_{s^n} \quad \forall n \in \mathbb{N}\right) \geq 1 - \varepsilon \qquad \forall s \in S,$$

where θ^n denotes the random variable for the history up to stage n and s^n the random variable for the state at stage n.

Note that although ε has two different roles in this definition, it will lead to no confusion.
The main reason for dealing with almost stationary ε-equilibria is that they can usually be handled easier than general ε-equilibria due to the simple structure of stationary strategies. However, they are also more appealing, as the players

settle on playing simple stationary strategies and only need to switch to other mixed actions with probabilities close to zero.
It is an interesting fact that, in zero-sum stochastic games, on the one hand ε-equilibria are known to exist for all $\varepsilon > 0$ (as we have mentioned in section 2.10, any pair of ε-optimal strategies yields a 2ε-equilibrium), but on the other hand the existence of almost stationary ε-equilibria has remained an open problem. In this chapter, which is based on Flesch et al. [1998,II], we will answer this question in the affirmative, namely we will construct almost stationary ε-equilibria, for all $\varepsilon > 0$, in all zero-sum stochastic games.

Main Theorem 6 *In every zero-sum stochastic game, there exists an almost stationary ε-equilibrium for any $\varepsilon > 0$.*

It is clear by lemma 2.9.7-(d),(e) that, in zero-sum stochastic games, 0-equilibria do not necessarily exist and that history dependent strategies are indispensable for obtaining ε-equilibria, $\varepsilon > 0$, so the result is sharp in this sense. Take a zero-sum stochastic game and let $\varepsilon > 0$. The proof will be based on a construction for a specific stationary strategy pair (x^*, y^*) with reward equal to the value. The value as a reward is acceptable for both players, as neither player is able to guarantee a better reward in his favor. So with the help of the stationary strategy pair (x^*, y^*), we will construct an almost stationary ε-equilibrium in which the players always use the mixed actions corresponding to the strategies x^* and y^*, unless a player detects that his opponent has used different mixed actions in the past. In order to detect such deviations, as a standard tool, the players will apply statistical tests on the past action frequencies of their opponents. If a player detects a deviation with a large certainty, then he has to start playing a history dependent δ-optimal strategy, where $\delta > 0$ is sufficiently small. The role of these δ-optimal strategies is to rule out the profitability of possible deviations of the players. To illustrate the issue, we will now briefly discuss the Big Match.

Example 6.1.2

	L	R
T	0	1
B	1 *	0 *

1

This game is the Big Match (cf. example 2.9.6). In lemma 2.9.7-(a),(b),(c) we discussed that the value for initial state 1 is $v_1 = 1/2$, and we presented an ε-optimal strategy π^ε for player 1 as well as the stationary optimal strategy $y = (1/2, 1/2)$ for player 2. This pair of strategies (π^ε, y) would be an ε-equilibrium that yields precisely $v_1 = 1/2$ to player 1. Instead of achieving this $1/2$ through this complicated strategy π^ε, the players could play the almost stationary ε-equilibrium (ξ^ε, y), where ξ^ε is the strategy defined, roughly speaking, by: play action T unless at some stage in the far future you notice that player 2's action frequencies are not sufficiently close to $(1/2, 1/2)$, in that case start playing π^ε immediately. Notice that if player 2 truly plays y, then the probability that player 2's action frequencies are not close enough to $(1/2, 1/2)$ in the far future is very small (by the law of large numbers). Hence, with probability almost 1, the players play stationary strategies forever.

6.2 Preliminaries

For $s \in S$, $x_s \in X_s$, $y_s \in Y_s$ let $L_s(x_s, y_s)$ be the probability that, after transition from state s with respect to (x_s, y_s), the new value v_t is different from v_s, so

$$L_s(x_s, y_s) := \sum_{t \in S,\, v_t \neq v_s} p_s(t|x_s, y_s)$$

(if $v_t = v_s$ for all $t \in S$ then $L_s(x_s, y_s)$ is defined to equal 0). Obviously, $V_s(x_s, y_s) \neq v_s$ implies $L_s(x_s, y_s) > 0$ (recall definition 3.2.3). The next lemma states that, with respect to (x, y), if the value does not change in expectation under transitions then the value is a constant on each ergodic set.

Lemma 6.2.1 *Let $(x, y) \in X \times Y$ satisfy $V(x, y) = v$. Suppose that E is an ergodic set with respect to (x, y). Then $v_s = v_t$ for all $s, t \in E$, and therefore $L_s(x_s, y_s) = 0$ for all $s \in E$.*

Proof. Let $\bar{E} := \{s \in E | \, v_s = \max_{t \in E} v_t\}$. By using $V(x, y) = v$ and the fact that E is an ergodic set for (x, y), we obtain

$$v_s = V_s(x_s, y_s) = \sum_{t \in S} p_s(t|x_s, y_s)\, v_t = \sum_{t \in E} p_s(t|x_s, y_s)\, v_t \qquad \forall s \in \bar{E},$$

thus $\bar{E} \subset E$ is a closed set of states for (x, y). Hence $\bar{E} = E$, so $v_s = v_t$ for all $s, t \in E$. Now $L_s(x_s, y_s) = 0$ for all $s \in E$ follows from the definition. $\square$

For $s \in S$ let

$$X'_s := \{x_s \in X_s | \, V_s(x_s, y_s) \geq v_s \quad \forall y_s \in Y_s\}, \qquad X' := \times_{s \in S} X'_s,$$

$$Y'_s := \{y_s \in Y_s | \, V_s(x_s, y_s) \leq v_s \quad \forall x_s \in X_s\}, \qquad Y' := \times_{s \in S} Y'_s,$$

$$\bar{X}_s := \{x_s \in X_s | \, V_s(x_s, y_s) = v_s \quad \forall y_s \in Y'_s\}, \qquad \bar{X} := \times_{s \in S} \bar{X}_s,$$

$$\bar{Y}_s := \{y_s \in Y_s | \, V_s(x_s, y_s) = v_s \quad \forall x_s \in X'_s\}, \qquad \bar{Y} := \times_{s \in S} \bar{Y}_s.$$

By lemmas 3.2.4 and 3.2.2, the above sets are nonempty polytopes. Let $\bar{I}_s$ and $\bar{J}_s$ denote the sets of extreme points of $\bar{X}_s$ and $\bar{Y}_s$, respectively. Recall that the relative interior of a polytope Z, denoted by $\text{Relint}(Z)$, is defined as the set of points in Z which can be written as a convex combination of all the extreme points of Z with only strictly positive coefficients. Due to lemma 3.2.2 again, for all $s \in S$, $x_s \in \text{Relint}(X'_s)$, $y_s \in \text{Relint}(Y'_s)$, we have

$$\bar{I}_s = \{i_s \in I_s | \, x_s(i_s) > 0\}, \qquad \bar{J}_s = \{j_s \in J_s | \, y_s(j_s) > 0\}. \tag{6.1}$$

The next lemma provides sufficient conditions for $\bar{X}_s = X'_s$ and for $\bar{Y}_s = Y'_s$ in some state $s \in S$.

Lemma 6.2.2 *Let $s \in S$. If $L_s(x_s, j_s) > 0$ implies $V_s(x_s, j_s) > v_s$ for all $(x_s, j_s) \in X'_s \times J_s$, then $\bar{X}_s = X'_s$. Similarly, if $L_s(i_s, y_s) > 0$ implies $V_s(i_s, y_s) < v_s$ for all $(i_s, y_s) \in I_s \times Y'_s$, then $\bar{Y}_s = Y'_s$.*

Proof. We will only show the first part; the proof of the second part is similar. By the definitions we have $\bar{X}_s \supset X'_s$, so it remains to verify that $\bar{X}_s \subset X'_s$. Since $\bar{X}_s$ and X'_s are convex, it is sufficient to show that $i_s \in X'_s$ for all $i_s \in \bar{I}_s$. Take an arbitrary $i_s \in \bar{I}_s$. Using the compactness of X'_s, there exists an $\hat{x}_s \in X'_s$ satisfying $\hat{x}_s(i_s) \geq x_s(i_s)$ for all $x_s \in X'_s$. By (6.1) we have $\hat{x}_s(i_s) > 0$.
For $\lambda \in (0, 1)$, let

$$x_s^\lambda := ((1 - \lambda) \cdot \hat{x}_s + \lambda \cdot i_s) \in X_s.$$

We will now show that $x_s^\lambda \in X'_s$ for small $\lambda > 0$. Since $V_s(x_s^\lambda, \cdot)$ is linear on Y_s and J_s is finite, we only need to show that, for any $j_s \in J_s$, if $\lambda > 0$ is small then $V_s(x_s^\lambda, j_s) \geq v_s$. Take an arbitrary $j_s \in J_s$. If $L_s(x_s^\lambda, j_s) = 0$ then $V_s(x_s^\lambda, j_s) = v_s$, so assume that $L_s(x_s^\lambda, j_s) > 0$. Then $\hat{x}_s(i_s) > 0$ implies $L_s(\hat{x}_s, j_s) > 0$, so by the condition we have $V_s(\hat{x}_s, j_s) > v_s$. Therefore, using the linearity of $V_s(\cdot, j_s)$ on X_s, we obtain $V_s(x_s^\lambda, j_s) \geq v_s$ if $\lambda > 0$ is small.

Hence $x_s^\lambda \in X_s'$ for small $\lambda > 0$ indeed. Now the choice of $\hat{x}_s$ yields $\hat{x}_s = i_s$, thus $i_s \in X_s'$. So the proof is complete. □

Thuijsman & Vrieze [1993] showed that, in every zero-sum game, there exists an initial state s_1 in $S^{\max} := \{s \in S | v_s = \max_{t \in S} v_t\}$ for which player 1 has a stationary optimal strategy x^1, and similarly, there exists an initial state s_2 in $S^{\min} := \{s \in S | v_s = \min_{t \in S} v_t\}$ for which player 2 has a stationary optimal strategy y^2. In view of theorem 2.8.2-(b), we may take a stationary best reply y^1 against x^1 and a stationary best reply x^2 against y^2. Let E^1 be an ergodic set with respect to (x^1, y^1) such that E^1 is entered with a positive probability, if the initial state is s_1 and the players use (x^1, y^1). As x^1 is optimal for s_1 and y^1 is a best reply, one can show by using lemma 2.7.1-(c) that $E^1 \subset S^{\max}$ and $\gamma_s(x^1, y^1) = v_s$ for all $s \in E^1$. By choosing an ergodic set E^2 for (x^2, y^2) in a similar way, we obtain the following result.

Lemma 6.2.3 *There exist stationary strategy pairs* (x^1, y^1), (x^2, y^2) *and corresponding ergodic sets* E^1, E^2 *such that*

$$E^1 \subset S^{\max} := \{s \in S | v_s = \max_{t \in S} v_t\}, \qquad \gamma_s(x^1, y^1) = v_s \quad \forall s \in E^1,$$

$$E^2 \subset S^{\min} := \{s \in S | v_s = \min_{t \in S} v_t\}, \qquad \gamma_s(x^2, y^2) = v_s \quad \forall s \in E^2.$$

6.3 The construction

Fix arbitrary $x' \in \text{Relint}(X')$ and $y' \in \text{Relint}(Y')$. We keep x' and y' fixed for the rest of this chapter. Let T denote the set of transient states and $\mathcal{R}$ the set of ergodic sets with respect to (x', y'). Since any stationary strategy pair induces at least one ergodic set, we have $\mathcal{R} \neq \emptyset$. Now we divide $\mathcal{R}$ into three parts. Let

$$\mathcal{R}^1 := \{E \in \mathcal{R} | \exists s \in E, \ \exists (i_s, y_s) \in I_s \times Y_s' : V_s(i_s, y_s) = v_s, \ L_s(i_s, y_s) > 0\}$$

$$\mathcal{R}^2 := \{E \in \mathcal{R} \setminus \mathcal{R}^1 | \exists s \in E, \ \exists (x_s, j_s) \in X_s' \times J_s : V_s(x_s, j_s) = v_s, \ L_s(x_s, j_s) > 0\}$$

$$\mathcal{R}^3 := \mathcal{R} \setminus (\mathcal{R}^1 \cup \mathcal{R}^2).$$

Note that all the sets $T, \mathcal{R}^1, \mathcal{R}^2, \mathcal{R}^3$ are independent of the particular choices of $x' \in \text{Relint}(X')$ and $y' \in \text{Relint}(Y')$, as all stationary strategies in $\text{Relint}(X')$ and $\text{Relint}(Y')$ put positive probabilities on the same actions in each state. Here $\mathcal{R}^1$ is the set of ergodic sets E with respect to (x', y') with the following property: there exists a pair of mixed actions in some state $s \in E$ such that player 1 plays a "pure" action, player 2 plays a mixed action in Y'_s, and the expected value after transition equals the original value, but with a positive probability a transition occurs to a state where the value is different. The intuition behind $\mathcal{R}^2$ is analogous. The partition of $\mathcal{R}$ naturally induces the following partition of $S \setminus T$:

$$S^1 := \cup_{E \in \mathcal{R}^1} E, \qquad S^2 := \cup_{E \in \mathcal{R}^2} E, \qquad S^3 := \cup_{E \in \mathcal{R}^3} E.$$

If $\mathcal{R}^1 \cup \mathcal{R}^2 \neq \emptyset$, then by the definitions of $\mathcal{R}^1$ and $\mathcal{R}^2$ there exists a nonempty set $S^* \subset S^1 \cup S^2$ with the following properties: (i) S^* contains precisely one state from each ergodic set in $\mathcal{R}^1 \cup \mathcal{R}^2$, (ii) for all $s \in S^* \cap S^1$, there exists a pair $(i^*_s, y^*_s) \in I_s \times Y'_s$ satisfying $V_s(i^*_s, y^*_s) = v_s$, $L_s(i^*_s, y^*_s) > 0$, (iii) for all $s \in S^* \cap S^2$, there exists a pair $(x^*_s, j^*_s) \in X'_s \times J_s$ satisfying $V_s(x^*_s, j^*_s) = v_s$, $L_s(x^*_s, j^*_s) > 0$. In fact, these states and pairs of mixed actions provide the possibility to leave all the ergodic sets belonging to $\mathcal{R}^1$ and $\mathcal{R}^2$ in such a way that the value does not change in expectation.

We will now turn our attention to $\mathcal{R}^3$. Assume that $E \in \mathcal{R}^3$ (in fact, later we will show that $\mathcal{R}^3$ is always nonempty). Since E is ergodic for (x', y'), by lemma 6.2.1 we have $v_s = v_t =: v_E$ for all $s, t \in E$. By using $E \cap (S^1 \cup S^2) = \emptyset$, in light of lemma 6.2.2, we have $\bar{X}_s = X'_s$ and $\bar{Y}_s = Y'_s$ for all $s \in E$. We may then define a restricted game $\bar{\Gamma}_E$ in which the state space is E and the players are restricted to using actions in $\bar{I}_s$ and $\bar{J}_s$, if the play is in any state $s \in E$. Obviously, this restricted game $\bar{\Gamma}_E$ is a well-defined stochastic game as well. In this restricted game, the respective stationary strategy spaces of the players are $\bar{X}_E := \times_{s \in E} \bar{X}_s$ and $\bar{Y}_E := \times_{s \in E} \bar{Y}_s$. We use $\bar{v}_s$, $s \in E$, for the value of the restricted game $\bar{\Gamma}_E$.

In the restricted game $\bar{\Gamma}_E$, in order to avoid confusion, we use $\bar{q}$ and $\bar{\gamma}$ instead of q and γ. In the next important lemma, we show the existence of stationary strategy pairs in $\bar{\Gamma}_E$ with rewards equal to the original value v_E.

Lemma 6.3.1 *Let $E \in \mathcal{R}^3$ and define the restricted game $\bar{\Gamma}_E$ as above. Then in $\bar{\Gamma}_E$, there exists a stationary strategy pair $(\bar{x}, \bar{y}) \in \bar{X}_E \times \bar{Y}_E$ such that $\bar{\gamma}_s(\bar{x}, \bar{y}) = v_E$ for all $s \in E$.*

Proof. We distinguish two essentially different cases.

Part 1: Assume that $\bar{v}_s \geq v_E$ for all $s \in E$ (if $\bar{v}_s \leq v_E$ for all $s \in E$, then an analogous proof can be applied). It follows from the first concluding remark in section 3.5 that there exists a state $s \in E$ such that $\bar{v}_s = v_E$.
Let $E^{\min} := \{t \in E | \, \bar{v}_t = v_E\}$. Let $(x^2, y^2) \in \bar{X}_E \times \bar{Y}_E$ and let $E^2 \subset E^{\min}$ as in lemma 6.2.3 for $\bar{\Gamma}_E$. So we have $\bar{\gamma}_s(x^2, y^2) = v_E$ for all $s \in E^2$. For $s \in E$ let

$$\bar{x}_s := \begin{cases} x_s^2 & \text{if } s \in E^2 \\ x'_s & \text{if } s \in E \setminus E^2 \end{cases}, \qquad \bar{y}_s := \begin{cases} y_s^2 & \text{if } s \in E^2 \\ y'_s & \text{if } s \in E \setminus E^2. \end{cases}$$

The only ergodic set for $(\bar{x}, \bar{y}) \in \bar{X}_E \times \bar{Y}_E$ in the restricted game $\bar{\Gamma}_E$ is E^2. Hence for any $s, t \in E$ we have that $\bar{q}_s(t|\bar{x}, \bar{y}) > 0$ only holds if $t \in E^2$, thus lemma 2.7.1-(c) yields $\bar{\gamma}_s(\bar{x}, \bar{y}) = v_E$ for all $s \in E$. $\diamond$

Part 2: Assume that $\min_{s \in E} \bar{v}_s < v_E < \max_{s \in E} \bar{v}_s$.
Take $(x^1, y^1) \in \bar{X}_E \times \bar{Y}_E$, $E^1 \subset E^{\max} := \{s \in E | \, \bar{v}_s = \max_{t \in E} \bar{v}_t\}$ and $(x^2, y^2) \in \bar{X}_E \times \bar{Y}_E$, $E^2 \subset E^{\min} := \{s \in E | \, \bar{v}_s = \min_{t \in E} \bar{v}_t\}$ in $\bar{\Gamma}_E$ as in lemma 6.2.3. By the assumption we have $E^1 \cap E^2 = \emptyset$. For $a, b \in (0,1)$ and $s \in E$ let

$$(x_s^{ab}, y_s^{ab}) := \begin{cases} (a \cdot x_s^1 + (1-a) \cdot x'_s, a \cdot y_s^1 + (1-a) \cdot y'_s) & \text{if } s \in E^1 \\ (b \cdot x_s^2 + (1-b) \cdot x'_s, b \cdot y_s^2 + (1-b) \cdot y'_s) & \text{if } s \in E^2 \\ (x'_s, y'_s) & \text{otherwise.} \end{cases}$$

Notice that $x_s^{ab} \in \text{Relint}(\bar{X}_s)$ and $y_s^{ab} \in \text{Relint}(\bar{Y}_s)$ for all $s \in E$ and $a, b \in (0,1)$, hence the set E is ergodic for (x^{ab}, y^{ab}) for all $a, b \in (0,1)$. Notice also that a and b control the respective expected lengths of periods when staying in E^1 and E^2. Since E is ergodic for (x^{ab}, y^{ab}) for all $a, b \in (0,1)$, lemma 2.7.1-(d) implies that $\bar{\gamma}_s(x^{ab}, y^{ab}) = \bar{\gamma}_t(x^{ab}, y^{ab}) := \bar{\gamma}_E^{ab}$ for all $s, t \in E$, $a, b \in (0,1)$.
It suffices to show that there are $a, b \in (0,1)$ such that $\bar{\gamma}_E^{ab} = v_E$. Take arbitrary $a', b' \in (0,1)$. If $\bar{\gamma}_E^{a'b'} = v_E$ then we are done. So assume without loss of generality that $\bar{\gamma}_E^{a'b'} > v_E$ and consider $(x^{a'b}, y^{a'b})$. Observe that the larger b we take, the more time the play spends in E^2. Thus one can show that

$$\lim_{b \uparrow 1} \bar{\gamma}_E^{a'b} = \min_{t \in E} \bar{v}_t < v_E.$$

By lemma 2.7.2, we have that $\bar{\gamma}_E^{a'b}$ is continuous in $b \in (0,1)$, hence there is a b such that $\bar{\gamma}_E^{a'b} = v_E$. $\square$

Now we are ready to complete the construction based on the previously derived results. Recall that we have already fixed a pair of stationary strategies $(x', y') \in \text{Relint}(X') \times \text{Relint}(Y')$. For all ergodic sets $E \in \mathcal{R}^3$ let $(\bar{x}_s, \bar{y}_s) \in \bar{X}_s \times \bar{Y}_s$, $s \in E$, be as in lemma 6.3.1. We define a stationary strategy pair (x^τ, y^τ) for all $\tau \in (0,1)$: for $s \in S$ let

$$(x_s^\tau, y_s^\tau) := \begin{cases} (\tau \cdot x'_s + (1-\tau) \cdot i_s^*, y_s^*) & \text{if } s \in S^* \cap S^1 \\ (x_s^*, \tau \cdot y'_s + (1-\tau) \cdot j_s^*) & \text{if } s \in S^* \cap S^2 \\ (\bar{x}_s, \bar{y}_s) & \text{if } s \in S^3 \\ (x'_s, y'_s) & \text{if } s \in S \setminus (S^* \cup S^3). \end{cases}$$

The next lemma shows that, for these stationary strategy pairs, the recurrent states all belong to S^3 and the reward equals the value for all initial states. This also implies that $S^3 \neq \emptyset$, therefore $\mathcal{R}^3 \neq \emptyset$.

Lemma 6.3.2 *Let $\tau \in (0,1)$. Then $\gamma(x^\tau, y^\tau) = v$. Furthermore, if U is an ergodic set with respect to (x^τ, y^τ), then $U \subset S^3$.*

Proof. Let $\tau \in (0,1)$. By the definitions, we have $V(x^\tau, y^\tau) = v$. Let U be an ergodic set with respect to (x^τ, y^τ). Then in view of lemma 6.2.1, $L_s(x_s^\tau, y_s^\tau) = 0$ for all $s \in U$. Hence the construction of (x^τ, y^τ) yields $U \subset S^3$.
The equality $V(x^\tau, y^\tau) = v$ implies $P(x^\tau, y^\tau)v = v$. By using induction, we have for all $n \in \mathbb{N}$ that $P^n(x^\tau, y^\tau)v = v$, hence the definition of $Q(x^\tau, y^\tau)$ yields

$$Q(x^\tau, y^\tau)\, v = v.$$

For any $s \in S$, if $q_s(t|x^\tau, y^\tau) > 0$ then t belongs to an ergodic set with respect to (x^τ, y^τ), so we have $t \in S^3$. Now the choice of $(\bar{x}_z, \bar{y}_z)$, $z \in S^3$, implies by lemma 6.3.1 that $\gamma_t(x^\tau, y^\tau) = v_t$ for all $t \in S^3$, so applying lemma 2.7.1-(c) gives

$$\begin{aligned} \gamma_s(x^\tau, y^\tau) &= \sum_{t \in S} q_s(t|x^\tau, y^\tau) \cdot \gamma_t(x^\tau, y^\tau) \\ &= \sum_{t \in S^3} q_s(t|x^\tau, y^\tau) \cdot v_t \\ &= v_s \qquad \forall s \in S, \end{aligned}$$

which completes the proof. □

Finally, we provide a proof for Main Theorem 6. We show that, for any $\varepsilon > 0$, the stationary strategy pair (x^τ, y^τ), for sufficiently large $\tau \in (0,1)$, can be

supplemented with history dependent δ-optimal strategies, for small $\delta > 0$, in order to obtain an almost stationary ε-equilibrium.

Proof of Main Theorem 6:
We only give an outline of the proof, since the tools used are standard (see for example Vrieze & Thuijsman [1989], Vieille [1993] or, in a more general fashion, Thuijsman & Vrieze [1996]). Let $\varepsilon > 0$. We will define strategy pairs (π^τ, σ^τ) for all $\tau \in (0,1)$ so that (π^τ, σ^τ) is an almost stationary ε-equilibrium for sufficiently large $\tau \in (0,1)$. These strategy pairs will be constructed in such a way that if neither player deviates, then the mixed actions according to the stationary strategy pair (x^τ, y^τ) are played forever with probability at least τ. In view of lemma 6.3.2, this means that if τ is large then

$$v_s - \frac{\varepsilon}{2} \leq \gamma_s(\pi^\tau, \sigma^\tau) \leq v_s + \frac{\varepsilon}{2} \qquad \forall s \in S.$$

Hence, when verifying the ε-equilibrium conditions, it suffices to show that

$$\gamma_s(\bar{\pi}, \sigma^\tau) \leq v_s + \frac{\varepsilon}{2} \qquad \forall s \in S,\ \forall \bar{\pi} \in \Pi$$

$$\gamma_s(\pi^\tau, \bar{\sigma}) \geq v_s - \frac{\varepsilon}{2} \qquad \forall s \in S,\ \forall \bar{\sigma} \in \Sigma.$$

The strategies π^τ and σ^τ will be analogously defined, so we only focus on player 1's strategy π^τ and on the possible deviations of player 2.
Let $\tau \in (0,1)$ be close to 1. Player 1's strategy π^τ will use the mixed actions according to x^τ unless, on condition that player 2 should play the mixed actions according to y^τ, player 1 detects with probability almost 1 that player 2 has deviated from y^τ. If player 1 detects such a deviation then player 1 starts playing a $(1-\tau)$-optimal strategy. Player 2's possible deviations are detected by means of employing statistical tests on player 2's behavior during the past history. Such a statistical test, with respect to some arbitrary stationary strategy pair $(\tilde{x}, \tilde{y})$, is based on the observations that, if player 2 truly uses his stationary strategy $\tilde{y}$, then: (1) player 2 never chooses an action which has probability zero with regard to $\tilde{y}$; (2) if the play remains in the same ergodic set (ergodic with respect to $(\tilde{x}, \tilde{y})$), then the empirical action frequencies of player 2 should converge to the weights of the mixed actions corresponding to $\tilde{y}$ (by the law of large numbers); (3) from any transient state (transient with respect to $(\tilde{x}, \tilde{y})$), the probability of remaining in the transient states longer than n stages converges to zero as n tends to infinity. So, if player 2 chooses an action with probability zero according to $\tilde{y}$, then player 1 will know for sure

that player 2 has deviated; if after some specified number of stages within an ergodic set, player 2's action frequencies are not within some specified range from the theoretical ones, then player 1 will suspect that player 2 has deviated; if the play remains in the set of transient states longer than some specified number of stages, then player 1 will suspect that player 2 has deviated.
In the discussion below we will assume that player 1 has not detected a deviation from player 2 yet, so player 1 still uses x^τ.

First we consider the case when player 2 chooses an action $j_s \in J_s$ in state $s \in S$ with $y_s^\tau(j_s) = 0$ (see observation (1) above). Then, clearly, player 1 immediately notices the deviation. So, using the inequalities $\lim_{\tau\uparrow 1} V_s(x_s^\tau, j_s) \geq v_s$ for all $j_s \in J_s$, $s \in S$, and the finiteness of the state and action spaces, if player 2 chooses any action $j_s \in J_s$ in any state $s \in S$ with $y_s^\tau(j_s) = 0$, then the reward is at least $V_s(x_s^\tau, j_s) - (1-\tau) \geq v_s - \frac{\varepsilon}{2}$ if τ is large enough; recall that π^τ prescribes a $(1-\tau)$-optimal strategy after such a deviation.

Now we assume the other case, namely that player 2 only prescribes actions which have positive probabilities with respect to y^τ. We divide the set of stages up to the current stage into blocks B^k of consecutive stages as follows: a new block starts at each stage the play enters T, or a set $E \in \mathcal{R}$, or an ergodic set U with respect to (x^τ, y^τ) (we must have $U \subset S^3$ in view of lemma 6.3.2). In block B^k the probability that, although player 2 truly used y^τ, player 1 detects a deviation of player 2 will be at most d^k, where $d^k \in (0,1)$ for all $k \in \mathbb{N}$ and $\sum_{k=1}^{\infty} d^k \leq 1-\tau$. The latter inequality will guarantee that the total probability of making this mistake is at most $1-\tau$.
Assume that the play enters some ergodic set $U \subset S^3$ (with respect to (x^τ, y^τ)) and the new block is B^k. In this ergodic set, player 1 checks the action frequencies of player 2, and if the empirical action frequencies are not close enough to the theoretical ones then player 1 detects a deviation (see observation (2) above with $(\tilde{x}, \tilde{y}) = (x^\tau, y^\tau)$) and starts playing a $(1-\tau)$-optimal strategy. If the number of stages in this block B^k is large enough, then the probability that player 1 detects a deviation although player 2 used y^τ is at most d^k. Note that if the empirical action frequencies are close to the theoretical ones, then the corresponding reward is close to the value. Notice that the play never leaves U if the players only use actions which are chosen with positive probabilities with respect to the pair (x^τ, y^τ).
Assume that the play enters T, or a set $E^2 \in \mathcal{R}^2$, or a set $E^3 \in \mathcal{R}^3$ but not an ergodic set $U \subset S^3$ (ergodic with respect to (x^τ, y^τ)), and that the new block is B^k. Then, by lemma 6.3.2, if player 2 uses y^τ, the play should leave T, or E^2, or enter an ergodic set $U \subset E^3 \subset S^3$, $U \neq E^3$ (ergodic with

respect to (x^τ, y^τ)) within N^k stages, for large N^k, with probability at least $1 - d^k$ (see observation (3) above with $(\tilde{x}, \tilde{y}) = (x^\tau, y^\tau)$). If this does not happen within N^k stages, then player 1 suspects player 2 of having deviated and starts playing a $(1-\tau)$-optimal strategy afterwards. Notice that $x_s^\tau \in X_s'$ for all $s \in T \cup S^2 \cup S^3$, hence the play can only leave in such a way that the value does not decrease in expectation.

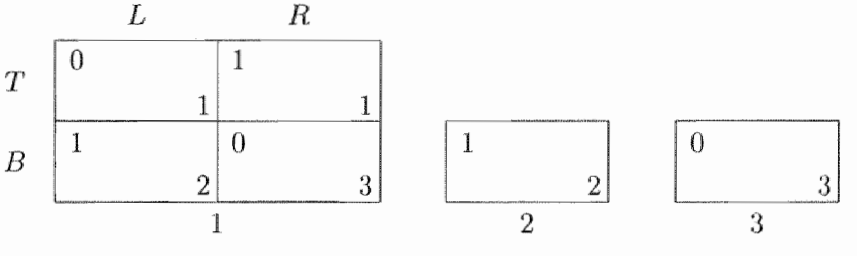
Finally, assume that the play enters some $E \in \mathcal{R}^1$ and the new block is B^k. Notice that, if τ is large, then x^τ almost equals x' in all states in E, thus the set E is "almost" ergodic for (x^τ, y^τ). Therefore, player 1 has enough time to check the action frequencies of player 2 in E (see observation (2) above with $(\tilde{x}, \tilde{y}) = (x', y^\tau)$). This way player 1 can make sure that the unique state s in $S^* \cap E$ is visited frequently enough and also that the play leaves E via i_s^* and the new value does not differ "much" from v_s (recall that $V_s(i_s^*, y_s^*) = v_s$). If player 2 truly uses y^τ then player 1 does not detect any deviation with probability at least $1 - d^k$.

We have described how player 1 makes sure that the reward is not much less than the value once the play reaches an ergodic set $U \subset S^3$ (ergodic with respect to (x^τ, y^τ)), and also that the play eventually reaches such an ergodic set in such a way that the value does not drop "much" in expectation. So if we take a sufficiently large $\tau \in (0,1)$, the proof is complete. □

6.4 Examples

We provide two examples to illustrate the construction of almost stationary ε-equilibria.

Example 6.4.1



We reexamine the Big Match, which we have also examined in example 6.1.2.

This example shows how the ergodic sets in $\mathcal{R}^1$ and $\mathcal{R}^2$ can be left in such a way that the value does not change "much" in expectation. In view of lemma 2.9.7-(a), the value is known to be $v = (1/2, 1, 0)$. Following the construction above, we have

$$X_1' = \{(1,0)\}, \quad X_2' = X_3' = \{(1)\},$$

$$Y_1' = \text{conv}\,\{(1/2,1/2),(0,1)\}, \quad Y_2' = Y_3' = \{(1)\},$$

$$\mathcal{R}^1 = \{\{1\}\}, \quad S^1 = \{1\}, \quad \mathcal{R}^2 = \emptyset, \quad S^2 = \emptyset,$$

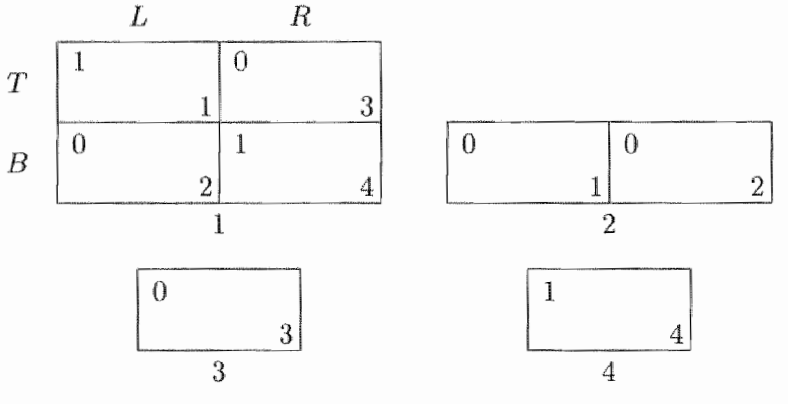
$$\mathcal{R}^3 = \{\{2\},\{3\}\}, \quad S^3 = \{2,3\},$$

where conv stands for the convex hull of a set. To see that $S^1 = \{1\}$ take $S^* = \{1\}$, $i_1^* = B$, $y_1^* = (1/2, 1/2)$. As X' is a singleton and states 2 and 3 are trivial, for $\tau \in (0,1)$ we have

$$x^\tau = ((\tau, 1-\tau),(1),(1)), \qquad y^\tau = ((1/2,1/2),(1),(1)).$$

Clearly, $\gamma(x^\tau, y^\tau) = v$ for all $\tau \in (0,1)$. Note that player 1 has no incentive to deviate from x^τ when playing against y^τ, because any strategy of player 1 would give reward 1/2 against y^τ, $\tau \in (0,1)$. On the other hand, if τ is large then player 1 is able to check the action frequencies of player 2 in state 1 with a high precision. Thus if the initial state is state 1 then player 1 can make sure that the eventual transitions to state 2 and state 3 will have almost equal probabilities, and then player 2 cannot gain more than an arbitrarily small ε by any deviation from y^τ.

Example 6.4.2



This example clarifies the above construction for ergodic sets in $\mathcal{R}^3$. Notice that states 3 and 4 are trivial. The value of the game is $v = (0, 0, 0, 1)$. To see that $v_1 = v_2 = 0$, take the stationary strategy $y^\delta = ((1-\delta, \delta), (0,1), (1), (1))$ for player 2, where $\delta \in (0,1)$. One can easily check that $\gamma_1(i, y^\delta) \leq \delta$ and $\gamma_2(i, y^\delta) = 0$ for all $i \in I$, so using theorem 2.8.2-(b) and the fact that the smallest payoff in the game is zero, we have $v_1 = v_2 = 0$ indeed. Note that the strategy y^δ is δ-optimal for player 2.
Following the construction above, we have

$$X' = X, \quad Y_1' = \{(1,0)\}, \quad Y_s' = Y_s \quad \forall s = 2, 3, 4,$$

$$\mathcal{R}^1 = \emptyset, \quad S^1 = \emptyset, \quad \mathcal{R}^2 = \emptyset, \quad S^2 = \emptyset,$$

$$\mathcal{R}^3 = \{\{1,2\}, \{3\}, \{4\}\}, \quad S^3 = \{1, 2, 3, 4\}.$$

We only focus on the ergodic set $E = \{1, 2\}$, as states 3 and 4 are trivial. Consider the restricted game $\bar{\Gamma}_E$. Let $\bar{v}_s$, $s = 1, 2$, denote the value of $\bar{\Gamma}_E$. Clearly,

$$\bar{v}_1 = 1 > 0 = v_1, \qquad \bar{v}_2 = 0 = v_2.$$

Now the strategies

$$\bar{x} = ((1/2, 1/2), (1)) \in \bar{X}_E, \quad \bar{y} = ((1,0), (0,1)) \in \bar{Y}_E$$

satisfy $\bar{\gamma}_s(\bar{x}, \bar{y}) = v_s$ for all $s \in E$ (cf. part 1 of the proof of lemma 6.3.1). So, for all $\tau \in (0,1)$ we have

$$x^\tau = ((1/2, 1/2), (1), (1), (1)), \qquad y^\tau = ((1,0), (0,1), (1), (1))).$$

We discuss how the pairs (x^τ, y^τ) can be used to obtain an almost stationary ε-equilibrium for some $\varepsilon > 0$. Clearly, $\gamma(x^\tau, y^\tau) = v$ for all $\tau \in (0,1)$. Note that player 2 has no incentive to deviate from y^τ when playing against x^τ. On the other hand, player 2 needs to check the action frequencies of player 1 in state 1, because player 1 could get reward 1 by playing action T at each stage, when playing against y^τ from initial state 1. So if the play does not leave state 1 after a large number of stages, then player 2 will suspect player 1 of having deviated and will start using the δ-optimal strategy y^δ, where δ is small. This assures that player 1 cannot improve his reward by more than ε. Note that the probability that player 1 truly uses x^τ, $\tau \in (0,1)$, but accidentally chooses action T for a very long time is small.

Part II

General-sum stochastic games

Chapter 7

Recursive repeated games with absorbing states

7.1 Introduction

In this chapter, which is based on Flesch et al. [1996], we deal with stochastic games in which all the states but one are absorbing, and in the non-absorbing state all the payoffs are equal to zero. Since these games are precisely those games which belong to the classes of recursive games and repeated games with absorbing states (cf. definition 2.11.1-(g),(h)), we call them recursive repeated games with absorbing states.

The main result, which will follow from theorem 7.3.1, is the following one.

Main Theorem 7 *In every recursive repeated game with absorbing states, there exists a stationary ε-equilibrium for any $\varepsilon > 0$.*

Several examples prove the sharpness of the result. Examples 7.3.4 and 2.10.3 demonstrate that stationary ε-equilibria, for small $\varepsilon > 0$, do not necessarily exist in recursive games and in repeated games with absorbing states. Moreover, example 7.3.3 shows that recursive repeated games with absorbing states do not always have stationary 0-equilibria. Finally, example 9.1.1 in chapter 9 will clarify why the above result fails to extend to games of this kind with more than two players.

7.2 Preliminaries

Consider a recursive repeated game with absorbing states and suppose that the initial state is an absorbing state $s \in S$. Then, by ignoring all the other states, we might as well consider the restricted game which only contains this single state s. In this restricted game, one can easily show that the average reward coincides with the β-discounted rewards, $\beta \in (0,1)$, for stationary strategies. Hence by using theorems 2.10.2 and 2.8.2-(b), stationary equilibria exist in this restricted game for the average reward. So, if we fix such a stationary equilibrium in state s, we may replace state s by another absorbing state with only one cell in which the payoffs equal the rewards corresponding to this stationary equilibrium of the original state. Therefore, without loss of generality, we assume that the absorbing states are of size 1×1.
Suppose that the non-absorbing state is state 1 and has size $m \times n$. As state 1 is the only non-trivial state, we suppress state 1 in the notations. Let $I := \{1, \ldots, m\}$ and $J := \{1, \ldots, n\}$ denote the action spaces of the players in state 1. Then the stationary strategy spaces X and Y have the form:

$$X = \left\{ x = x(i)_{i \in I} | \sum_{i \in I} x(i) = 1, \ x(i) \geq 0 \quad \forall i \in I \right\}$$

$$Y = \left\{ y = y(j)_{j \in J} | \sum_{j \in J} y(j) = 1, \ y(j) \geq 0 \quad \forall j \in J \right\}.$$

Assume that the initial state is state 1. Notice that, as state 1 is non-absorbing, there has to be another state as well. Since the game only contains one non-absorbing state by the conditions, it means that the other state must be absorbing. Hence there is at least one absorbing state in the game.
If entry $(i,j) \in I \times J$ of state 1 is chosen, then with probability

$$p^*_{ij} := \sum_{s \in S \setminus \{1\}} p(s|i,j) = 1 - p(1|i,j)$$

absorption occurs in the set of absorbing states with expected absorption payoff a^k_{ij} for player $k \in \{1,2\}$, and with probability $1 - p^*_{ij}$ the play stays in the initial state with payoffs zero. For completeness we define $a^k_{ij} := 0$ for $k = 1,2$ if $p^*_{ij} = 0$.

Definition 7.2.1 *Let $x \in X$ and $y \in Y$. Let*

$$p^*_{xy} := \sum_{i\in I} \sum_{j\in J} x(i)\, y(j)\, p^*_{ij},$$

$$T^1(y) := \{i \in I \mid p^*_{iy} > 0\},$$

$$B^1(y) := \{i \in I \mid \gamma^1(i,y) \geq \gamma^1(\pi,y) \quad \forall \pi \in \Pi\}.$$

The sets $T^2(x)$ and $B^2(x)$ are analogously defined. For the strategy pair (x,y) we have that p^*_{xy} is the one step absorption probability. If $p^*_{xy} > 0$ then (x,y) eventually leads to absorption, while if $p^*_{xy} = 0$ then absorption never occurs with regard to (x,y). Therefore, in the former case we call (x,y) absorbing, while in the latter case we say that (x,y) is non-absorbing. Now $T^1(y)$ consists of the pure stationary strategies (or actions) that are absorbing against y, and $B^1(y)$ is the set of pure stationary best replies against y. Similar definitions and interpretations apply for $T^2(x)$ and $B^2(x)$. In view of theorem 2.8.2-(b), the sets $B^1(y)$ and $B^2(x)$ are always non-empty.
The next lemma provides explicit expressions for the reward when stationary strategies x and y are used. If (x,y) is non-absorbing then the reward is 0, while if (x,y) is absorbing then the reward equals a convex combination of the rewards for (i,y), $i \in I$, where the coefficients are precisely the probabilities that the absorption occurs when player 1 plays actions $i \in I$.

Lemma 7.2.2 *Let $(x,y) \in X \times Y$ and $k \in \{1,2\}$. If $p^*_{xy} = 0$ then $\gamma^k(x,y) = 0$, while if $p^*_{xy} > 0$ then*

$$\gamma^k(x,y) = \frac{\sum_{i\in I} \sum_{j\in J} x(i)\, y(j)\, p^*_{ij}\, a^k_{ij}}{p^*_{xy}}$$

and

$$\gamma^k(x,y) = \frac{\sum_{i\in I} x(i)\, p^*_{iy}\, \gamma^k(i,y)}{\sum_{i\in I} x(i)\, p^*_{iy}} = \frac{\sum_{i\in T^1(y)} x(i)\, p^*_{iy}\, \gamma^k(i,y)}{\sum_{i\in T^1(y)} x(i)\, p^*_{iy}}.$$

Proof. Let $(x,y) \in X \times Y$ and $k \in \{1,2\}$. If $p^*_{xy} = 0$ then the play remains in state 1 forever with probability 1. As all the payoffs in state 1 equal zero, we obtain $\gamma^k(x,y) = 0$.
Assume now that $p^*_{xy} > 0$. As $p^*_{xy} > 0$, the pair (x,y) is absorbing, thus $\gamma^k(x,y)$ equals the expected absorption payoff:

$$\gamma^k(x,y) = \frac{\sum_{i\in I} \sum_{j\in J} x(i)\ y(j)\, p^*_{ij}\, a^k_{ij}}{p^*_{xy}}$$

(this expression also follows from lemma 2.7.1-(c)). Similarly, if $p^*_{iy} > 0$ for some $i \in I$ then

$$\gamma^k(i,y) = \frac{\sum_{j\in J}\ y(j)\, p^*_{ij}\, a^k_{ij}}{p^*_{iy}}.$$

Hence

$$\begin{aligned}\frac{\sum_{i\in I}\ x(i)\, p^*_{iy}\, \gamma^k(i,y)}{\sum_{i\in I}\ x(i)\, p^*_{iy}} &= \frac{\sum_{i\in I}\ x(i)\ \sum_{j\in J}\ y(j)\, p^*_{ij}\, a^k_{ij}}{\sum_{i\in I}\ x(i)\ \sum_{j\in J}\ y(j)\, p^*_{ij}} \\ &= \frac{\sum_{i\in I}\ \sum_{j\in J}\ x(i)\ y(j)\, p^*_{ij}\, a^k_{ij}}{p^*_{xy}} \\ &= \gamma^k(x,y).\end{aligned}$$

Since $p^*_{xy} > 0$, we must have $T^1(y) \neq \emptyset$. Thus the definition of $T^1(y)$ implies

$$\gamma^k(x,y) = \frac{\sum_{i\in T^1(y)}\ x(i)\, p^*_{iy}\, \gamma^k(i,y)}{\sum_{i\in T^1(y)}\ x(i)\, p^*_{iy}},$$

so the proof is complete. □

Next, we introduce proper and δ-proper strategy pairs, where $\delta \in (0,1)$.

Definition 7.2.3 *A pair of stationary strategies $(x_\delta, y_\delta) \in X \times Y$ is called δ-proper for $\delta > 0$, if*

(a) *(x_δ, y_δ) is completely mixed, namely $x_\delta(i) > 0$ for all $i \in I$ and $y_\delta(j) > 0$ for all $j \in J$,*

(b) *$\gamma^1(i, y_\delta) > \gamma^1(i', y_\delta)$ implies $\delta \cdot x_\delta(i) \geq x_\delta(i')$ for all $i, i' \in I$,*

(c) *$\gamma^2(x_\delta, j) > \gamma^2(x_\delta, j')$ implies $\delta \cdot y_\delta(j) \geq y_\delta(j')$ for all $j, j' \in J$.*

A pair of stationary strategies $(x,y) \in X \times Y$ is called proper, if

$$(x,y) = \lim_{n\to\infty} (x_n, y_n)$$

for some sequence of δ_n-proper strategy pairs (x_n, y_n), where δ_n is a positive and monotonously decreasing sequence converging to 0.

The notions of proper and δ-proper strategy pairs are very similar to those of so-called proper and δ-proper equilibria in normal form games (cf. Myerson [1978] and van Damme [1991]). However, here proper and δ-proper strategy pairs do not necessarily correspond to (ε-)equilibria, for small $\varepsilon > 0$, as shown in the following example.

Example 7.2.4

	L	R
T	0,0	2,−2 *
B	1,−1 *	0,0 *

1

In this game, entry (T, L) is non-absorbing and all other entries are absorbing with probability 1. Note that although, formally, all the payoffs in state 1 should equal to zero, it makes no difference for the average reward whether the payoffs in the absorbing cells differ from zero. Here $((1-\delta^2, \delta^2), (1-\delta^2, \delta^2))$ is δ-proper for small $\delta > 0$, so $((1,0),(1,0))$ is proper, but neither one is an ε-equilibrium for small $\varepsilon > 0$. One can argue as follows. The pair $((1,0),(1,0))$ is clearly no ε-equilibrium for $\varepsilon \in [0,1)$, since player 1 can improve his reward by 1 if he chooses action B. On the other hand, for any $\delta \in (0,1)$, the pair $((1-\delta^2, \delta^2), (1-\delta^2, \delta^2))$ is no ε-equilibrium for $\varepsilon \in [0, 1/2]$ either, because by using lemma 7.2.2

$$\begin{aligned} \gamma^1((1-\delta^2,\delta^2),(1-\delta^2,\delta^2)) &= \frac{(1-\delta^2)\delta^2 \cdot 2 + \delta^2\left[(1-\delta^2)\cdot 1 + \delta^2 \cdot 0\right]}{(1-\delta^2)\delta^2 + \delta^2} \\ &= \frac{3(1-\delta^2)}{2-\delta^2} \\ &< \frac{3}{2}, \end{aligned}$$

while action T gives 2 against $(1-\delta^2, \delta^2)$. ◁

Theorem 7.2.5 *In any recursive repeated game with absorbing states, there exists a proper strategy pair.*

Proof. As $X \times Y$ is compact, any sequence in $X \times Y$ must have a convergent subsequence. Therefore it suffices to show that, for sufficiently small $\delta > 0$,

there exists a δ-proper pair. Recall that $|I| = m$ and $|J| = n$. Choose a sufficiently small $\delta > 0$ so that the following sets are nonempty for all $x \in X$ and $y \in Y$:

$$X_\delta := \{x \in X \mid x(i) \geq \delta^m \quad \forall i \in I\}$$

$$Y_\delta := \{y \in Y \mid y(j) \geq \delta^n \quad \forall j \in J\}$$

$$X_\delta(y) := \{x \in X_\delta | \gamma^1(i,y) > \gamma^1(i',y) \text{ implies } \delta \cdot x(i) \geq x(i') \ \forall i, i' \in I\}$$

$$Y_\delta(x) := \{y \in Y_\delta | \gamma^2(x,j) > \gamma^2(x,j') \text{ implies } \delta \cdot y(j) \geq y(j') \ \forall j, j' \in J\}.$$

Note that the sets X_δ, Y_δ, and $X_\delta(y)$, $Y_\delta(x)$ are polytopes. Consider the following correspondence Ψ from $X_\delta \times Y_\delta$ to the set of all subsets of $X_\delta \times Y_\delta$:

$$\Psi(x,y) = (X_\delta(y), Y_\delta(x)).$$

By lemma 7.2.2 (or by lemma 2.7.2), $\gamma^1(i,\cdot)$ is continuous on Y_δ for all $i \in I$, and $\gamma^2(\cdot,j)$ is continuous on X_δ for all $j \in J$. Therefore, the correspondence Ψ has the property (upper hemi-continuity) that: if $w^n \to w$ and $\bar{w}^n \to \bar{w}$ in $X_\delta \times Y_\delta$ as $n \to \infty$, and $w^n \in \Psi(\bar{w}^n)$ for all $n \in \mathbb{N}$, then we necessarily have $w \in \Psi(\bar{w})$. By Kakutani's fixed point theorem (cf. Kakutani [1941]), Ψ must have a fixed point, namely a $w = (x,y) \in X_\delta \times Y_\delta$ such that $w \in \Psi(w)$, or equivalently, $(x,y) \in X_\delta(y) \times Y_\delta(x)$. Because every fixed point is a δ-proper pair, the proof is complete. □

7.3 The construction

Let $(x_\delta, y_\delta) \in X \times Y$ for all $\delta > 0$. By the finiteness of the action spaces I and J, there must exist a countable subset of $\mathcal{D}$ of positive real numbers such that 0 is a limit point of $\mathcal{D}$; the sets $\{i \in I | x_\delta(i) > 0\}$, $\{j \in J | y_\delta(j) > 0\}$, $T^1(y_\delta)$, $T^2(x_\delta)$, $B^1(y_\delta)$, and $B^2(x_\delta)$ are independent of $\delta \in \mathcal{D}$. By using compactness arguments, we may furthermore assume that the following sequences have limits as δ tends to 0 in $\mathcal{D}$: (i) $x_\delta(i^1)/x_\delta(i^2)$ for all $i^1, i^2 \in I$, whenever $x_\delta(i^2) > 0$; (ii) $y_\delta(j^1)/y_\delta(j^2)$ for all $j^1, j^2 \in J$, whenever $y_\delta(j^2) > 0$; (iii) (x_δ, y_δ). In this chapter, each time that we are dealing with limits when δ converges to zero, we will have such a subset $\mathcal{D}$ in mind.

The following theorem provides a construction for stationary ε-equilibria for all $\varepsilon > 0$ in recursive repeated games with absorbing states.

Theorem 7.3.1 *In a recursive repeated game with absorbing states, let* $(\tilde{x}, \tilde{y})$ *be a proper pair with*

$$(\tilde{x}, \tilde{y}) = \lim_{\delta \downarrow 0, \delta \in \mathcal{D}} (x_\delta, y_\delta),$$

where (x_δ, y_δ) *is* δ*-proper for all* $\delta \in \mathcal{D}$*. Then for any* $\varepsilon > 0$

(a) *if* $(\tilde{x}, \tilde{y})$ *is absorbing, then* (x_δ, y_δ) *is an* ε*-equilibrium for small* $\delta \in \mathcal{D}$*,*

(b) *if* $(\tilde{x}, \tilde{y})$ *is non-absorbing, then* $(\tilde{x}, \tilde{y})$ *or* $(x_\delta, \tilde{y})$ *or* $(\tilde{x}, y_\delta)$ *is an* ε*-equilibrium for small* $\delta \in \mathcal{D}$*.*

The following example provides an illustration for the above construction.

Example 7.3.2

	L	R
T	0,0	4,−3 *
M	3,−2 *	1,−4 *
B	1,−4 *	3,−2 *

1

Here entry (T, L) is non-absorbing while all the other entries lead to absorption with probability 1. Recall that, although formally all the payoffs in state 1 should equal to zero, it makes no difference for the average reward whether the payoffs in the absorbing cells differ from zero.
We will now illustrate the above construction for stationary ε-equilibria, where $\varepsilon > 0$. Notice that the stationary strategy pair

$$(x_\delta, y_\delta) = ((1 - \delta^2 - \delta^4, \delta^4, \delta^2), (\delta^2, 1 - \delta^2))$$

is δ-proper for small $\delta > 0$, hence

$$(\tilde{x}, \tilde{y}) = ((1, 0, 0), (0, 1))$$

is proper. Here $(\tilde{x}, \tilde{y})$ is absorbing, and one can easily check that, for any $\varepsilon > 0$, the stationary strategy pair (x_δ, y_δ) is an ε-equilibrium for small $\delta > 0$.

Note that $(\tilde{x}, \tilde{y})$ is not an ε-equilibrium in this game for small $\varepsilon > 0$, because player 2 would be better off by choosing action L with probability 1.
The pair

$$(x_\delta, y_\delta) = ((1 - \delta^2 - \delta^4, \delta^2, \delta^4), (1 - \delta^2, \delta^2))$$

is also δ-proper for small $\delta > 0$, so

$$(\tilde{x}, \tilde{y}) = ((1, 0, 0), (1, 0))$$

is proper. Here $(\tilde{x}, \tilde{y})$ is non-absorbing, and action M of player 1 is a profitable best reply against $\tilde{y}$ and leads to absorption in entry (M, L). In order to make player 1 satisfied, we let player 1 play x_δ with a sufficiently small $\delta > 0$. Observe that, for small $\delta > 0$, the pair $(x_\delta, \tilde{y})$ also leads to absorption in entry (M, L) with probability close to 1. So for any $\varepsilon > 0$, the strategy x_δ is an ε-best reply against $\tilde{y}$ if $\delta > 0$ is small. On the other hand, $\tilde{y}$ is obviously a best reply against x_δ. Hence for any $\varepsilon > 0$, the stationary strategy pair $(x_\delta, \tilde{y})$ is an ε-equilibrium for small $\delta > 0$ indeed. ◁

The next example demonstrates that stationary equilibria need not necessarily exist in recursive repeated games with absorbing states.

Example 7.3.3

	L	R
T	0,0	1,−1 *
B	1,−1 *	0,0 *

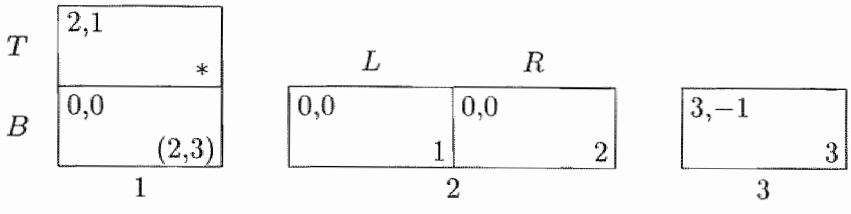
1

We will show that there are no stationary equilibria in this game. Recall that although, formally, all the payoffs in state 1 should equal to zero, it makes no difference for the average reward whether the payoffs in the absorbing cells differ from zero.
Assume by way of contradiction that $(x, y) \in X \times Y$ is an equilibrium. Then if y puts a positive probability on action R then x has to choose action T with probability 1, which contradicts the fact that y does not play action L with probability 1. On the other hand, if y takes action L with probability 1, then x must choose action B with a positive probability, which is in contradiction

with the fact that y does not choose action R with probability 1. So we may conclude that no stationary equilibrium exists in this game. ◁

With the help of the following example, we illustrate that Main Theorem 7 does not extend to recursive games (cf. definition 2.11.1-(h)).

Example 7.3.4



Here $(2,3)$ stands for the transition vector which brings the play to state 2 with probability $1/2$ and to state 3 with probability $1/2$. Recall that although all the payoffs in state 1 should equal to zero, it makes no difference for the average reward whether the payoffs in the absorbing cells differ from zero.
We will now prove that there are no stationary equilibria in this game; one can similarly show that this game does not posses stationary ε-equilibria for small $\varepsilon > 0$ either. We represent stationary strategies of the players by the probabilities on actions T and L, respectively. Suppose by way of contradiction that (x, y) is an equilibrium. If $y > 0$ then $x = 0$ must hold, which contradicts $y > 0$ from player 2's point of view. On the other hand, if $y = 0$ then we must have $x = 1$, which is in contradiction with $y = 0$ from player 2's point of view. Hence there are no stationary equilibria in this game indeed. ◁

7.4 The proof

The first lemma deals with stationary ε-best replies, $\varepsilon > 0$, of the players.

Lemma 7.4.1 *Let $\varepsilon > 0$. Let $(x_\delta, y_\delta) \in X \times Y$ for all $\delta \in \mathcal{D}$, and let $\tilde{y} := \lim_{\delta \downarrow 0, \delta \in \mathcal{D}} y_\delta$. Suppose that there exists an action $i^* \in B^1(y_\delta) \cap T^1(\tilde{y})$ such that $x_\delta(i^*) > 0$. If*

$$\lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \frac{x_\delta(i)}{x_\delta(i^*)} = 0 \qquad \forall i \in T^1(y_\delta) \setminus B^1(y_\delta),$$

then, for sufficiently small $\delta \in \mathcal{D}$, the strategy x_δ is an ε-best reply against y_δ. A similar statement holds for player 2 as well.

Proof. We only show the statement for player 1. Notice that $i^* \in T^1(\tilde{y})$ implies

$$0 < p^*_{i^* y} = \lim_{\delta \downarrow 0,\, \delta \in \mathcal{D}} p^*_{i^* y_\delta},$$

hence $i^* \in T^1(y_\delta)$. So lemma 7.2.2 yields that for sufficiently small $\delta \in \mathcal{D}$ we have

$$\begin{aligned}
\gamma^1(x_\delta, y_\delta) &= \frac{\sum_{i \in T^1(y_\delta)} x_\delta(i)\, p^*_{iy_\delta}\, \gamma^1(i, y_\delta)}{\sum_{i \in T^1(y_\delta)} x_\delta(i)\, p^*_{iy_\delta}} \\
&= \frac{\sum_{\substack{i \in T^1(y_\delta) \\ i \in B^1(y_\delta)}} x_\delta(i)\, p^*_{iy_\delta}\, \gamma^1(i, y_\delta) + \sum_{\substack{i \in T^1(y_\delta) \\ i \notin B^1(y_\delta)}} x_\delta(i)\, p^*_{iy_\delta}\, \gamma^1(i, y_\delta)}{\sum_{\substack{i \in T^1(y_\delta) \\ i \in B^1(y_\delta)}} x_\delta(i)\, p^*_{iy_\delta} + \sum_{\substack{i \in T^1(y_\delta) \\ i \notin B^1(y_\delta)}} x_\delta(i)\, p^*_{iy_\delta}} \\
&= \frac{\sum_{\substack{i \in T^1(y_\delta) \\ i \in B^1(y_\delta)}} \frac{x_\delta(i)}{x_\delta(i^*)}\, p^*_{iy_\delta}\, \gamma^1(i, y_\delta) + \sum_{\substack{i \in T^1(y_\delta) \\ i \notin B^1(y_\delta)}} \frac{x_\delta(i)}{x_\delta(i^*)}\, p^*_{iy_\delta}\, \gamma^1(i, y_\delta)}{\sum_{\substack{i \in T^1(y_\delta) \\ i \in B^1(y_\delta)}} \frac{x_\delta(i)}{x_\delta(i^*)}\, p^*_{iy_\delta} + \sum_{\substack{i \in T^1(y_\delta) \\ i \notin B^1(y_\delta)}} \frac{x_\delta(i)}{x_\delta(i^*)}\, p^*_{iy_\delta}} \\
&\geq \frac{\sum_{\substack{i \in T^1(y_\delta) \\ i \in B^1(y_\delta)}} \frac{x_\delta(i)}{x_\delta(i^*)}\, p^*_{iy_\delta}\, \gamma^1(i, y_\delta)}{\sum_{\substack{i \in T^1(y_\delta) \\ i \in B^1(y_\delta)}} \frac{x_\delta(i)}{x_\delta(i^*)}\, p^*_{iy_\delta}} - \varepsilon \\
&= \gamma^1(i^*, y_\delta) - \varepsilon.
\end{aligned}$$

As $i^* \in B^1(y_\delta)$, the proof is complete. □

Now we are ready to prove theorem 7.3.1.

Proof of theorem 7.3.1.
(a) Let $\varepsilon > 0$. Since $(\tilde{x}, \tilde{y})$ is absorbing, there exists an $i^* \in T^1(\tilde{y})$ such that $\tilde{x}(i^*) > 0$. Then the δ-properness of (x_δ, y_δ) implies that $i^* \in B^1(y_\delta)$ and also that

$$\lim_{\delta \downarrow 0,\, \delta \in \mathcal{D}} \frac{x_\delta(i)}{x_\delta(i^*)} = 0 \qquad \forall i \in T^1(y_\delta) \setminus B^1(y_\delta),$$

so the conditions of lemma 7.4.1 are fulfilled and therefore x_δ is an ε-best reply against y_δ for sufficiently small $\delta \in \mathcal{D}$. One can similarly show that y_δ is also an ε-best reply against x_δ for sufficiently small $\delta \in \mathcal{D}$. Therefore (x_δ, y_δ) is an ε-equilibrium on condition that $\delta \in \mathcal{D}$ is sufficiently small.

(b) Let $\varepsilon > 0$ and assume that $(\tilde{x}, \tilde{y})$ is non-absorbing. If $(\tilde{x}, \tilde{y})$ is an equilibrium, then we are done. Otherwise, at least one of the players has a profitable deviation with respect to $(\tilde{x}, \tilde{y})$. Suppose without loss of generality that player 1 has a profitable deviation. We will then show that $(x_\delta, \tilde{y})$ must be an ε-equilibrium for sufficiently small $\delta \in \mathcal{D}$.
Let $i^* \in B^1(\tilde{y})$ be a profitable best reply of player 1 against $\tilde{y}$. Then, since $(\tilde{x}, \tilde{y})$ is non-absorbing and i^* is a profitable deviation, we must have $i^* \in T^1(\tilde{y})$. Since $i^* \in T^1(\tilde{y})$, we also have $i^* \in T^1(y_\delta)$. Assume $i \in T^1(\tilde{y}) \setminus B^1(\tilde{y})$. Then $i \in T^1(y_\delta)$ as well and by lemma 2.7.2 (or by lemma 7.2.2)

$$\lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \gamma^1(i^*, y_\delta) = \gamma^1(i^*, \tilde{y}) > \gamma^1(i, \tilde{y}) = \lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \gamma^1(i, y_\delta).$$

Hence $\gamma^1(i^*, y_\delta) > \gamma^1(i, y_\delta)$ for sufficiently small $\delta \in \mathcal{D}$. By the δ-properness of (x_δ, y_δ) we have $x_\delta(i^*) > 0$ for all $\delta \in \mathcal{D}$ and

$$\lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \frac{x_\delta(i)}{x_\delta(i^*)} = 0.$$

So by lemma 7.4.1 with $(x_\delta, \tilde{y})$ instead of (x_δ, y_δ), we obtain that x_δ is an ε-best reply against $\tilde{y}$ for small $\delta \in \mathcal{D}$.
On the other hand, $\tilde{y}$ is a best reply against x_δ. One can argue as follows. Let $j^* \in J$ satisfy $\tilde{y}(j^*) > 0$. Then

$$\lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \frac{y_\delta(j^*)}{y_\delta(j)} > 0 \qquad \text{for all } j \in J, \tag{7.1}$$

so, by the δ-properness of (x_δ, y_δ), we must have $\gamma^2(x_\delta, j^*) \geq \gamma^2(x_\delta, j)$ for all $j \in J$ and for small $\delta \in \mathcal{D}$. Hence $j^* \in B^2(x_\delta)$ for all $j^* \in J$ with $\tilde{y}(j^*) > 0$. This yields in view of lemma 7.2.2 that the stationary strategy $\tilde{y}$ is also a best reply against x_δ.
Therefore we may conclude that $(x_\delta, \tilde{y})$ is an ε-equilibrium, if $\delta \in \mathcal{D}$ is sufficiently small. □

7.5 Concluding remarks

There is another way to establish equilibria having similar properties as δ-proper pairs by defining the following restricted strategy spaces for small $\delta > 0$

$$\bar{X}_\delta := \left\{ x \in X \mid \sum_{i \in U} x(i) \geq \delta^{|I|-|U|} \quad \forall\, \emptyset \neq U \subset I \right\}$$

$$\bar{Y}_\delta := \left\{ y \in Y \mid \sum_{j \in V} y(j) \geq \delta^{|J|-|V|} \quad \forall\, \emptyset \neq V \subset J \right\},$$

where $|Z|$ denotes the cardinality of a set Z, and by defining "linearized" rewards

$$\bar{\gamma}^1(x,y) := \sum_{i \in I} x(i)\, \gamma^1(i,y), \qquad \bar{\gamma}^2(x,y) := \sum_{j \in J} y(j)\, \gamma^2(x,j).$$

By applying Kakutani's fixed point theorem (cf. Kakutani [1941]), one can show the existence of stationary equilibria $(\bar{x}_\delta, \bar{y}_\delta)$ in $\bar{X}_\delta \times \bar{Y}_\delta$ with respect to the rewards $(\bar{\gamma}^1, \bar{\gamma}^2)$. Such equilibria have similar properties as δ-proper pairs, and the existence of stationary ε-equilibria can therefore be established analogously.

Notice that a stationary equilibrium (z_δ, w_δ) would also exist in $\bar{X}_\delta \times \bar{Y}_\delta$ with respect to the original rewards (γ^1, γ^2), but for such an equilibrium $\gamma^1(i, w_\delta) > \gamma^1(i', w_\delta)$ would not necessarily imply $\delta \cdot z_\delta(i) \geq z_\delta(i')$. This causes a discontinuity in the best reply structures when approaching $X \times Y$ by $\bar{X}_\delta \times \bar{Y}_\delta$.

Chapter 8

Average-discounted equilibria

8.1 Introduction

In this chapter, which is mainly based on Flesch et al. [1998,III], we investigate existence of equilibria in stochastic games in which the players use different evaluations. We assume that player 1 uses the average reward, while player 2 is interested in his β-discounted reward, $\beta \in (0,1)$. We will call these games average-discounted games. By the nature of these rewards, player 1 is interested in the far future payoffs, while player 2's reward is rather determined by the near future payoffs. So, different time periods interest the players, which may lead to a natural cooperation between them. First we define what we mean by equilibria in these games.

Definition 8.1.1 *A strategy pair (π, σ) is called an average-β-discounted ε-equilibrium, where $\beta \in (0,1)$ and $\varepsilon \geq 0$, if for all $s \in S$, $\bar{\pi} \in \Pi$, $\bar{\sigma} \in \Sigma$, we have*

$$\gamma_s^1(\bar{\pi}, \sigma) \leq \gamma_s^1(\pi, \sigma) + \varepsilon \quad \text{and} \quad \gamma_{\beta s}^2(\pi, \bar{\sigma}) \leq \gamma_{\beta s}^2(\pi, \sigma) + \varepsilon.$$

The main results of this chapter, which will follow from theorems 8.2.1 and 8.3.1, can be summarized as follows.

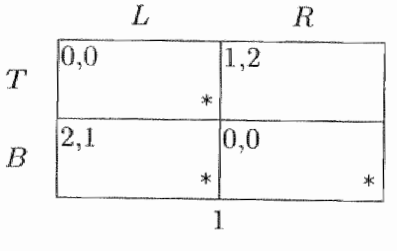
Main Theorem 8

(a) *In any stochastic game, for any $\beta \in (0,1)$ and $\varepsilon > 0$, there exists a stationary average-β-discounted ε-equilibrium.*

(b) *In any stochastic game, for any $\beta \in (0,1)$ and $\varepsilon > 0$, there exist Markov average-β-discounted ε-equilibria (f,g) such that there is a stage N with the following properties: $f_s(n) = f_s(N)$ and $g_s(n) = g_s(N)$ for any $s \in S$ and $n \geq N$, and if the play is in state s at stage N, then player 1 receives $\rho_s := \sup_{\pi \in \Pi, \sigma \in \Sigma} \gamma_s^1(\pi, \sigma)$.*

In view of (a), stationary strategies are sufficient for establishing ε-equilibria, $\varepsilon > 0$, in these average-discounted games. The existence of ε-equilibria in terms of stationary strategies is appealing, since stationary strategies are simple strategies. However, these stationary ε-equilibria have the draw-back that they do not make use of the special nature of these games, namely, they do not make use of the fact that different time periods interest the players. Therefore, in (b), we also prove the existence of ε-equilibria with the property that, after a sufficiently large stage when the discounted game is not interesting any longer, the players "cooperate" to guarantee the highest feasible reward to player 1. These ε-equilibria are formed by only slightly more complex Markov strategies, which we will call "ultimately stationary" since the strategies behave as if they were stationary strategies after finitely many stages.
Example 8.4.1 will demonstrate that average-β-discounted 0-equilibria do not always exist, not even in terms of history dependent strategies, so the result is sharp. We will also examine the existence of average-discounted ε-equilibria, $\varepsilon > 0$, in special classes of stochastic games.
We will now briefly discuss the following game to clarify the issues.

Example 8.1.2



Take an arbitrary discount factor $\beta \in (0,1)$. There are two really simple stationary average-β-discounted 0-equilibria. One of them is playing entry (B, L) at stage 1, yielding absorption with rewards $(2,1)$, while the other one is playing entry (T, R) at each stage, giving rewards $(1,2)$. These stationary equilibria, however, are not in the spirit of the game. The players should

decide to play entry (T, R) sufficiently long so that player 2's reward, which is rather determined by the near future payoffs, becomes almost 2, and then, when the rest of the play does not really interest player 2 any longer, to play entry (B, L) so as to give player 1 his highest feasible payoff (namely payoff 2) at each further stage. This plan, yielding rewards close to $(2, 2)$, can be realized by applying ultimately stationary strategies. Note that rewards close to $(2, 2)$ cannot be obtained by stationary average-β-discounted ε-equilibria, with small $\varepsilon > 0$. $\triangleleft$

8.2 Stationary ε-equilibria

This section is devoted to the existence of stationary ε-equilibria, $\varepsilon > 0$, in these average-discounted games. First we will introduce a restricted strategy space for player 2. Let

$$\bar{\delta} := \min_{s \in S} \frac{1}{|J_s|}.$$

For $\delta \in (0, \bar{\delta}]$ let

$$Y_s(\delta) := \{y_s \in Y_s \,|\, y_s(j_s) \geq \delta \quad \forall j_s \in J_s\} \quad \forall s \in S, \quad Y(\delta) := \times_{s \in S} Y_s(\delta);$$

in words, $Y(\delta)$ is the set of stationary strategies of player 2 which use each action in each state with probability at least δ. Obviously, $Y(\delta)$ is a polytope, and by the choice of $\bar{\delta}$ it is nonempty.
The main result of this section is the following theorem.

Theorem 8.2.1 *In any stochastic game, there exists a stationary average-β-discounted ε-equilibrium for any $\beta \in (0, 1)$ and $\varepsilon > 0$.*

Proof. Take arbitrary $\beta \in (0, 1)$ and $\varepsilon > 0$. By lemma 2.7.5, the function $\gamma^2_{\beta s}(\cdot, \cdot)$ is continuous on the compact space $X \times Y$, for any $s \in S$, hence it is uniformly continuous as well. Therefore there exists a $\delta \in (0, \bar{\delta}]$ such that for all $s \in S$ we have

$$\sup_{x \in X} \left[\sup_{y \in Y} \gamma^2_{\beta s}(x, y) - \sup_{y \in Y(\delta)} \gamma^2_{\beta s}(x, y) \right] \leq \varepsilon. \tag{8.1}$$

Consider the restricted game $\Gamma(\delta)$ which is derived from the original game Γ by restricting player 2 to using mixed actions in the set $Y_s(\delta)$ if the play is in any state $s \in S$.

It can be shown in the restricted game $\Gamma(\delta)$ analogously to theorem 2.8.4 that, against any fixed stationary strategy of player 1, there always exists a stationary β-discounted best reply y for player 2 such that y is an extreme point of $Y(\delta)$. It also holds similarly to theorem 2.10.2 that there exists a stationary (α, β)-discounted equilibrium (x^α, y^α) in $\Gamma(\delta)$, for all $\alpha \in (0,1)$. So for any $\alpha \in (0,1)$ we have

$$\gamma^1_{\alpha s}(x^\alpha, y^\alpha) \geq \gamma^1_{\alpha s}(i, y^\alpha) \quad \forall s \in S,\ \forall i \in I$$

$$\gamma^2_{\beta s}(x^\alpha, y^\alpha) \geq \gamma^2_{\beta s}(x^\alpha, \bar{y}) \quad \forall s \in S,\ \forall \bar{y} \in Y(\delta). \tag{8.2}$$

By the finiteness of the state and action spaces and by using compactness arguments, there exists a countable subset of discount factors $\mathcal{A} \subset (0,1)$ such that (i) 1 is a limit point of $\mathcal{A}$, (ii) the sets $\{i_s \in I_s | \, x^\alpha_s(i_s) > 0\}$, $\{j_s \in J_s | \, y^\alpha_s(j_s) > 0\}$, $s \in S$, are independent of $\alpha \in \mathcal{A}$, (iii) the sequence (x^α, y^α) has a limit in the compact space $X \times Y$ as α tends to 1 in $\mathcal{A}$. Let

$$(x, y) := \lim_{\alpha \uparrow 1,\ \alpha \in \mathcal{A}} (x^\alpha, y^\alpha)$$

We will now show that (x, y) is an average-β-discounted ε-equilibrium in the original game Γ. First we show that x is a best reply against y (for the average reward). As

$$x = \lim_{\alpha \uparrow 1,\ \alpha \in \mathcal{A}} x^\alpha,$$

we must have

$$\{i_s \in I_s | \, x_s(i_s) > 0\} \subset \{i_s \in I_s | \, x^\alpha_s(i_s) > 0\} \qquad \forall \alpha \in \mathcal{A}.$$

Therefore, by theorem 2.8.4, x is an α-discounted best reply against y^α for all $\alpha \in \mathcal{A}$, which means that

$$\gamma^1_{\alpha s}(x, y^\alpha) \geq \gamma^1_{\alpha s}(i, y^\alpha) \quad \forall s \in S,\ \forall i \in I,\ \forall \alpha \in \mathcal{A}.$$

As $y \in Y(\delta)$, applying lemma 2.7.6 yields for all $s \in S$ and $i \in I$ that

$$\gamma^1_s(x, y) = \lim_{\alpha \uparrow 1,\ \alpha \in \mathcal{A}} \gamma^1_{\alpha s}(x, y^\alpha) \geq \lim_{\alpha \uparrow 1,\ \alpha \in \mathcal{A}} \gamma^1_{\alpha s}(i, y^\alpha) = \gamma^1_s(i, y),$$

so in view of theorem 2.8.2-(b), the strategy x is a best reply against y (for the average reward).

We will now show that y is a β-discounted best reply against x in the original game Γ. By using lemma 2.7.5 and (8.2), we have for all $s \in S$ and $\bar{y} \in Y(\delta)$ that

$$\gamma^2_{\beta s}(x,y) = \lim_{\alpha\uparrow 1,\, \alpha\in\mathcal{A}} \gamma^2_{\beta s}(x^\alpha, y^\alpha) \geq \lim_{\alpha\uparrow 1,\, \alpha\in\mathcal{A}} \gamma^2_{\beta s}(x^\alpha, \bar{y}) = \gamma^2_{\beta s}(x,\bar{y}).$$

Therefore (8.1) implies for all $s \in S$

$$\gamma^2_{\beta s}(x,y) \geq \sup_{\bar{y}\in Y(\delta)} \gamma^2_{\beta s}(x,\bar{y}) \geq \sup_{\bar{y}\in Y} \gamma^2_{\beta s}(x,\bar{y}) - \varepsilon,$$

so, by theorem 2.8.4, y is a β-discounted best reply against x in the original game Γ. Hence, (x,y) is an average-β-discounted ε-equilibrium indeed. □

8.3 Ultimately stationary ε-equilibria

In this section, we will show the existence of ε-equilibria in terms of Markov strategies, which are ultimately stationary, meaning that the prescribed mixed actions become stage independent after finitely many stages. The idea is that, after a large stage N, player 2 becomes uninterested in the game due to the large powers of the discount factor β, so the players can "cooperate" to guarantee the highest feasible reward for player 1 in the future. During the first N stages, obviously, player 1 has to be careful not to ruin his future perspectives after stage N. We can state it formally as follows.

Theorem 8.3.1 *In any stochastic game, for any $\beta \in (0,1)$ and $\varepsilon > 0$, there exist Markov average-β-discounted ε-equilibria (f,g) such that there is a stage N with the following properties: $f_s(n) = f_s(N)$ and $g_s(n) = g_s(N)$ for any $s \in S$ and $n \geq N$, and if the play is in state s at stage N, then player 1 receives $\rho_s := \sup_{\pi\in\Pi,\sigma\in\Sigma} \gamma^1_s(\pi,\sigma)$.*

Proof. Take a stochastic game Γ. It is known that there exists a pure stationary strategy pair $(i^*,j^*) \in I \times J$ such that

$$\rho_s = \gamma^1_s(i^*,j^*) \qquad \forall s \in S,$$

where ρ_s is defined as in the theorem. (The existence of such a pure stationary strategy pair $(i^*,j^*) \in I \times J$ stems from the theory of Markov decision processes. In fact, such pairs $(i^*,j^*) \in I \times J$ are exactly the pure optimal solutions of the Markov decision process which is derived from the game by assuming

that there is only one player with action space $I_s \times J_s$, payoff function r_s^1, transition map p_s in states $s \in S$, and reward function γ^1).
Take arbitrary $\beta \in (0,1)$ and $\varepsilon > 0$. Let $K \in \mathbb{N}$, $K \geq 2$, be so large that

$$\beta^K \cdot \left[\max_{s,i_s,j_s} r_s^2(i_s,j_s) - \min_{s,i_s,j_s} r_s^2(i_s,j_s) \right] \leq \varepsilon,$$

which guarantees that, after stage K, player 2 can only improve his β-discounted reward by at most ε.
Consider the game Γ^K which is played until the play arrives at the state at stage $K+1$ and in which player 1 maximizes the expected value of $\rho_{s^{K+1}}$ (where s^{K+1} denotes the random variable for the state at stage $K+1$) and player 2 maximizes his K-stage β-discounted reward. Using backwards induction, one can construct a K-stage Markov average-β-discounted 0-equilibrium (f^K, g^K) in the game Γ^K in the following manner. We only discuss the construction briefly. For the final stage K when actions have to be chosen, let

$$(f_s^K(K), g_s^K(K)) \in X_s \times Y_s$$

be a 1-stage (Markov average-β-discounted) 0-equilibrium in any state $s \in S$. Given the mixed actions at stage K, we can choose

$$(f_s^K(K-1), g_s^K(K-1)) \in X_s \times Y_s$$

in states $s \in S$ so that

$$(f_s^K(m), g_s^K(m))_{s \in S, m = K-1, K}$$

is a 2-stage Markov average-β-discounted 0-equilibrium. By repeating this latter step inductively, finally we obtain a K-stage Markov average-β-discounted 0-equilibrium (f^K, g^K).
We define a pair of Markov strategies (f,g) as follows: for $s \in S$ and $n \in \mathbb{N}$ let

$$f_s(n) := \begin{cases} f_s^K(n) & \text{if } n \leq K \\ i_s^* & \text{if } n > K \end{cases} \qquad g_s(n) := \begin{cases} g_s^K(n) & \text{if } n \leq K \\ j_s^* & \text{if } n > K \end{cases}.$$

So f denotes the Markov strategy which coincides with f^K for the first K stages and which prescribes the pure stationary strategy i^* afterwards. The interpretation of g is analogous. Thus by their definitions, f and g satisfy the requirements of the theorem with $N := K+1$. We only have to show that (f,g) is an ε-equilibrium. Observe that player 1's average reward γ^1 is completely determined by the value of ρ in the state at stage $K+1$, which is

exactly what he maximizes in expectation during the first K stages, so player 1 cannot improve his reward at all. Player 2 cannot improve his reward during the first K stages, by his reward function in the game Γ^K, while after stage K, he can only improve it by ε because of the choice of K. So (f, g) is an average-β-discounted ε-equilibrium indeed. □

8.4 A game without 0-equilibria

In the previous two sections we showed the existence of ε-equilibria, $\varepsilon > 0$, in terms of stationary and ultimately stationary strategies. The following interesting example will demonstrate that, in these average-discounted games, 0-equilibria do not always exist, not even in history dependent strategies. So as it might be expected, the solutions of average-discounted games are on the one hand more complex than that of discounted games, where stationary 0-equilibria always exist (cf. theorem 2.10.2), but on the other hand simpler than that of average games, where stationary ε-equilibria do not generally exist for small $\varepsilon > 0$ (cf. example 2.10.3).

Example 8.4.1

	L	R
T	1,−1 *	−1,1 *
B	−1,1 *	0,0

1

Let $\beta \in (0, 1)$. We will show that, in the above game, there are no average-β-discounted 0-equilibria for initial state 1.
Notice that strategies only need to be defined for past histories where the initial state is 1 and no absorption has occurred. It is clear that, during such a past history, the players always chose actions B and R, hence the only information carried by such a history is the current stage. This means that all history dependent strategies are simply Markov strategies.
We may represent any mixed action in state 1 by the probability assigned to the first action. Therefore, any Markov strategy for any player is an element of the set $\times_{n=1}^{\infty}[0, 1]$.

Suppose by way of contradiction that the pair $(f,g) = (f(n),g(n))_{n=1}^{\infty}$ is a Markov 0-equilibrium with respect to $(\gamma^1,\gamma_\beta^2)$; here $f(n)$ and $g(n)$ denote the probabilities of playing action T and L, respectively, at stage n. Let $f^k := (f(n))_{n=k}^{\infty}$ and $g^k := (g(n))_{n=k}^{\infty}$ for any $k \in \mathbb{N}$, so f^k and g^k are the Markov strategies f and g starting from stage k. Let ξ^k denote player 1's average reward when using (f^k,g^k) for initial state 1.
Based on the assumption that (f,g) is a 0-equilibrium, we subsequently derive that we should have

(a) $\xi^1 > -1$;

(b) $0 < f(1) < 1$ and $0 < g(1) < 1$;

(c) (f^n,g^n) is a 0-equilibrium, $0 < f(n) < 1$, and $0 < g(n) < 1$ for all $n \in \mathbb{N}$;

(d) $\xi^n < \xi^{n+1}$ and $g(n) < g(n+1)$ for all $n \in \mathbb{N}$.

Afterwards we will show that these properties lead to a contradiction.

Proof of (a):
Since (f,g) is a 0-equilibrium, it suffices to define a Markov strategy $\bar{f}$ for player 1 which guarantees a reward larger than -1 when playing against g. For $n \in \mathbb{N}$ let

$$\bar{f}(n) := \begin{cases} 1 & \text{if } g(n) > 0 \\ 0 & \text{if } g(n) = 0. \end{cases}$$

Now with respect to $(\bar{f},g)$, whenever the play is in state 1, either the cell (B,R) is played with probability 1 or the cell (T,L) is played with a positive probability, hence $\gamma_1^1(\bar{f},g) > -1$.

Proof of (b):
If $f(1) = 1$ then $g(1) = 0$, since it yields absorption in entry (T,R) giving the highest possible reward for player 2. However, this contradicts (a), hence $f(1) < 1$ must hold.
If $f(1) = 0$ then $g(1) = 1$, which is also in contradiction with (a). Hence $f(1) > 0$ must be the case.
If $g(1) = 1$ then $f(1) = 1$ has to hold because f is a best reply against g. This, however, contradicts $0 < f(1) < 1$, so we must have $g(1) < 1$.
Now suppose that $g(1) = 0$. Using (a) we have

$$-1 < \xi^1 = f(1) \cdot (-1) + (1 - f(1)) \cdot \xi^2,$$

thus by $f(1) > 0$ we obtain $\xi^2 > \xi^1$. This means that player 1 would be better off by playing action B at stage 1 and playing f^2 from stage 2 on, which would assure reward ξ^2. This is in contradiction with the fact that $f(1) > 0$, so $g(1) > 0$ must hold.

Proof of (c):
By (b), the probability of no absorption at stage 1 has a positive probability. Therefore, clearly, (f^2, g^2) must be a 0-equilibrium as well. Using that (f^2, g^2) is a 0-equilibrium, one can show similarly to (b) that $0 < f(2) < 1$ and $0 < g(2) < 1$. Hence, the probability of no absorption at stage 2 has a positive probability too, so (f^3, g^3) must also be a 0-equilibrium. Now repeating this argument yields the statement.

Proof of (d):
The strategy f^1 is a best reply against g^1 and player 1 plays action B with a positive probability at stage 1 in view of (c), hence

$$\xi^1 = g(1) \cdot (-1) + (1 - g(1)) \cdot \xi^2.$$

Now by using (a) and $g(1) > 0$ (by (c)), we have $\xi^1 < \xi^2$ indeed. Repeating this argument leads to $\xi^n < \xi^{n+1}$ for all $n \in \mathbb{N}$.
Let $n \in \mathbb{N}$ be arbitrary. In view of (c), player 1 plays action T with positive probabilities at stages n and $n+1$, and f^n and f^{n+1} are best replies against g^n and g^{n+1} respectively, thus

$$\xi^n = g(n) \cdot 1 + (1 - g(n)) \cdot (-1)$$

$$\xi^{n+1} = g(n+1) \cdot 1 + (1 - g(n+1)) \cdot (-1).$$

Now it follows from $\xi^n < \xi^{n+1}$ that $g(n) < g(n+1)$.

Deriving a contradiction:
By (c), for any $m \in \mathbb{N}$, the strategy f^m is a best reply against g^m and player 1 plays action B with a positive probability at stage m, hence

$$\xi^m = g(m) \cdot (-1) + (1 - g(m)) \cdot \xi^{m+1}. \tag{8.3}$$

Consider the strategy $\bar{f}_K$, $K \geq 2$, which prescribes action B up to stage $K-1$ and the strategy f^K from stage K on. Then with respect to $(\bar{f}_K, g)$, player 1's reward is -1 if absorption occurs during the first $K-1$ stages and equals ξ^K otherwise. Thus, by applying (8.3), we have for all $K \geq 2$ that

$$\gamma^1(\bar{f}_K, g) =$$

$$= \left[1 - \prod_{n=1}^{K-1} (1 - g(n))\right] \cdot (-1) + \left[\prod_{n=1}^{K-1} (1 - g(n))\right] \cdot \xi^K$$

$$= \left[1 - \prod_{n=1}^{K-2} (1 - g(n))\right] \cdot (-1) +$$

$$+ \left[\prod_{n=1}^{K-2} (1 - g(n))\right] \cdot \left[g(K-1) \cdot (-1) + (1 - g(K-1)) \cdot \xi^K\right]$$

$$= \left[1 - \prod_{n=1}^{K-2} (1 - g(n))\right] \cdot (-1) + \left[\prod_{n=1}^{K-2} (1 - g(n))\right] \cdot \xi^{K-1}$$

$$= \cdots$$

$$= g(1) \cdot (-1) + (1 - g(1)) \cdot \xi^2$$

$$= \xi^1.$$

However, by properties (b) and (d) we have that $g(n) > g(1) > 0$ for all $n \in \mathbb{N}$. Therefore, if player 1 uses $\bar{f}_K$ with a large K then absorption occurs in entry (B, L) during the first $K - 1$ stages with probability almost 1. Formally,

$$\lim_{K \to \infty} \left[1 - \prod_{n=1}^{K-1} (1 - g(n))\right] = 1,$$

thus

$$\xi^1 = \lim_{K \to \infty} \gamma^1(\bar{f}_K, g) = -1,$$

which contradicts (a). Hence the basic assumption that (f, g) is a 0-equilibrium is false. □

8.5 Special classes of stochastic games

This section is devoted to a brief study of average-discounted equilibria in special classes of stochastic games, in which (ε-)equilibria can be achieved by using other techniques. Recall definition 2.11.1 for the precise definitions of the special classes considered below.

Unichain games. In unichain stochastic games there is just one ergodic set of states for any pair of stationary strategies, which assures that the average reward $\gamma_s^1(\cdot,\cdot)$ is also continuous on $X \times Y$ for all $s \in S$. In these games one can establish stationary average-β-discounted 0-equilibria, for any $\beta \in (0,1)$, as in the proof of theorem 8.2.1 without using restrictions on the stationary strategy space Y of player 2.

Perfect information and ARAT games. In perfect information games and ARAT games, one can establish average-β-discounted 0-equilibria, by using arguments as in Thuijsman & Raghavan [1997]. The idea is that player 1 has to play a pure stationary average optimal strategy i in his own game, namely an i with

$$\inf_{\sigma\in\Sigma} \gamma_s^1(i,\sigma) = \sup_{\pi\in\Pi} \inf_{\sigma\in\Sigma} \gamma_s^1(\pi,\sigma) \qquad \forall s \in S$$

(cf. theorem 2.11.2-(b),(d)), and player 2 has to play a stationary β-discounted best reply y against the strategy i (cf. theorem 2.8.4). This already implies that player 2 does not have a profitable deviation against i. Notice that, since the strategy i prescribes one pure action for each state, player 2 can immediately detect any deviation of player 1. Now in order to eliminate the profitability of deviations of player 1, if player 2 detects a deviation from i then he has to punish player 1 by switching to a strategy σ satisfying $\gamma_s^1(\pi,\sigma) \leq \gamma_s^1(i,y)$ for all s and π. Note that these punishments are effective due to the transition structure of these games.

Repeated games with absorbing states. In these games, one can establish average-β-discounted ε-equilibria, for all $\beta \in (0,1)$ and $\varepsilon > 0$, as follows. As stated in theorem 2.10.2, for any $\alpha \in (0,1)$, there exists a stationary equilibrium (x^α, y^α) with respect to $(\gamma_\alpha^1, \gamma_\beta^2)$; notice that we do not use restrictions on the strategy spaces at the moment. As in the proof of theorem 8.2.1, let (x,y) denote the limit strategy pair.

By using techniques as in Vrieze & Thuijsman [1989], one can show that either (x, y) or (x, y^α) with a large α can be supplemented with history dependent "punishment" strategies to establish an ε-equilibrium with respect to $(\gamma^1, \gamma^2_\beta)$.

8.6 Concluding remarks

We wish to remark that, in the literature of stochastic games and Markov decision processes, games have already been studied where, instead of using the discounted or the average evaluation, the players (or the player) use convex combinations of several discounted rewards with different discount factors and the average reward (cf. for example Filar & Vrieze [1992], Feinberg & Shwartz [1994], [1995]). Although the ideas have something in common, the arising problems require a different analysis.

Chapter 9

More than two players

9.1 Introduction

The model and the solution concepts of two-person general-sum stochastic games naturally extend to stochastic games with more than two players. However, the analysis becomes substantially more difficult, as demonstrated below. In two-person stochastic games, many of the usual techniques for establishing equilibria (just like the techniques in the previous two chapters) are based on sequences of stationary equilibria in auxiliary games that approach the original game in a certain sense. In these auxiliary games, mostly either the strategy spaces are specifically restricted or the reward functions are approximated by continuous functions (for example by the discounted rewards).
In this chapter, which is based on Flesch et al. [1997,I], we demonstrate that K-person stochastic games, with $K \geq 3$, require an analysis that is substantially different from any analysis used for two-person games. This is done by examining the following three-person stochastic game.

Example 9.1.1

In the following cubic three-person game, there is only one non-absorbing state, in which each player has two actions. The actions of the players are denoted as follows: player 1: T (top), B (bottom); player 2: L (left), R (right); player 3: N (near), F (far). The game is represented by taking separately the two layers of the cube that belong to the two actions of player 3 (N and F). Absorbing entries are indicated by $*'$s as usual.

Consider the game Γ :

N

	L	R
T	0,0,0	0,1,3 *
B	1,3,0 *	1,0,1 *

F

3,0,1 *	1,1,0 *
0,1,1 *	0,0,0 *

The most important properties of this game are summarized as follows.

Main Theorem 9 *The three-person stochastic game Γ in example 9.1.1 has the following properties.*

(a) *There are no stationary ε-equilibria for sufficiently small $\varepsilon > 0$.*

(b) *Let the Markov strategies κ, λ, μ for players 1,2,3 be respectively given by*

$$\kappa = \left(\frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, \ldots\right)$$

$$\lambda = \left(0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \ldots\right)$$

$$\mu = \left(0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, \ldots\right),$$

where the n-th coordinates of the above strategies are the probabilities for the second actions of the players at stage n. Then (κ, λ, μ) is a Markov equilibrium with equilibrium rewards $\gamma(\kappa, \lambda, \mu) = (1, 2, 1)$.

(c) *For any $\varepsilon \geq 0$, ε-equilibria only exist in terms of Markov strategies.*

(d) *Let (κ, λ, μ) be a Markov equilibrium with the property that, at any stage, at least one of the players plays his second action with a positive probability. Then, at each stage, exactly one of the players plays his second action with a positive probability, and these players appear cyclically in the order 1,2,3.*

(e) *The set of equilibrium rewards is the triangle (without its interior) in* $\mathbb{R}^3$ *with extreme points* $(1,1,2)$, $(1,2,1)$, $(2,1,1)$.

In the above theorem, property (c) is clarified at the beginning of section 9.2. The other properties will follow from theorems 9.2.3, 9.2.4, 9.2.6 and 9.2.7. Apparently, this is the first three-person stochastic game studied in detail. In fact, this game is a three-person recursive repeated game with absorbing states (as we discussed in chapter 7, it makes no difference for the average reward whether or not the payoffs in the absorbing cells differ from zero). In chapter 7, for two-person recursive repeated games with absorbing states, we showed the existence of stationary ε-equilibria for all $\varepsilon > 0$. The above example therefore demonstrates that the two-person result does not extend to stochastic games of this kind with more than two players.
The gap between two-person and three-person stochastic games also appears in the nature of equilibria. Supposedly, this is the first stochastic game where (cyclic) Markov strategies are indispensable, so the class of almost stationary strategy pairs (cf. definition 6.1.1) is too narrow to tackle the equilibrium existence problem for stochastic games with more than two players.

9.2 The analysis

Consider the game Γ in example 9.1.1. Since state 1 is the only non-absorbing state, we assume that the initial state is state 1. For simplicity, we will suppress state 1 in the notations. Note that all entries but entry (T, L, N) are absorbing, so the play absorbs as soon as one of the players chooses his second action. Notice that the payoff and the transition structure is cyclically symmetric, namely it holds for any entry $(i_1, i_2, i_3) \in \{1,2\}^3$ that

$$r^1(i_1,i_2,i_3) = r^2(i_2,i_3,i_1) = r^3(i_3,i_1,i_2)$$

$$p(i_1,i_2,i_3) = p(i_2,i_3,i_1) = p(i_3,i_1,i_2).$$

However we wish to emphasize that we have only introduced this cyclic symmetry to make the analysis of this game clearer. Similar results can also be obtained in non-symmetric games with the very same absorption structure.
In the game Γ, each mixed action can be represented by the probability assigned to the second action, which lets the stationary strategy spaces equal $[0,1]$ for each player. For stationary strategies of the players we use the notations x, y and z respectively.

The sets of Markov strategies equal $\times_{n=1}^{\infty}[0,1]$ for each player. Markov strategies for the players are denoted by κ, λ and μ respectively.
For this game the only history up to stage $n \in \mathbb{N}$, if no absorption has occurred, is the trivial one where all the players have chosen their first action at all stages up to stage n. Therefore all history dependent strategies are only Markov strategies. This observation implies Main Theorem 9-(c).
Now we investigate the game Γ in detail.

Lemma 9.2.1 *There is no stationary equilibrium in* Γ.

Proof. Suppose by way of contradiction that (x, y, z) is a stationary equilibrium. First we prove that $0 < x, y, z < 1$. Recall that x, y, z are the probabilities on actions B, R and F respectively. If $x = 0$ then, because of a best reply argument, $y = 1$ and therefore $z = 0$, which is in contradiction with $x = 0$. On the other hand $x = 1$ would imply $y = 0$, hence $z = 1$, which leads to a contradiction with $x = 1$. So $0 < x < 1$, and by symmetry we also have $0 < y, z < 1$.
Since $0 < x < 1$ and x is a best reply to (y, z), we must have (for instance, by applying similar expressions as in lemma 7.2.2) that

$$\frac{3(1-y)z + yz}{1-(1-y)(1-z)} = \gamma^1(0, y, z) = \gamma^1(1, y, z) = 1 - z,$$

thus

$$y = \frac{z^2 + 2z}{z^2 + 1} > z.$$

By symmetry $z > x$ and $x > y$. Hence $y > z > x > y$, contradiction. □

We call a triple (x, y, z) of stationary strategies in the game Γ absorbing, if $x > 0$ or $y > 0$ or $z > 0$. Such an absorbing triple eventually leads to absorption with probability 1. On the other hand, a triple (x, y, z) in the game Γ is called non-absorbing, if $x = y = z = 0$; in this case entry (T, L, N) is played forever with probability 1.

Lemma 9.2.2 *Let* (x_n, y_n, z_n) *be a sequence of stationary strategy triples in the game* Γ *with* $(\tilde{x}, \tilde{y}, \tilde{z}) := \lim_{n\to\infty}(x_n, y_n, z_n)$.

(a) *Assume that* $(\tilde{x}, \tilde{y}, \tilde{z})$ *is absorbing. Then*

$$\gamma(\tilde{x}, \tilde{y}, \tilde{z}) := \lim_{n\to\infty} \gamma(x_n, y_n, z_n).$$

(b) *Assume that* $(\tilde{x}, \tilde{y}, \tilde{z})$ *is non-absorbing,* (x_n, y_n, z_n) *are absorbing for all* $n \in \mathbb{N}$, *and the limits of the sequences*

$$w_{(x_n,y_n,z_n)}(B,L,N) := \frac{x_n\,(1-y_n)\,(1-z_n)}{1-(1-x_n)\,(1-y_n)\,(1-z_n)}$$

$$w_{(x_n,y_n,z_n)}(T,R,N) := \frac{(1-x_n)\,y_n\,(1-z_n)}{1-(1-x_n)\,(1-y_n)\,(1-z_n)}$$

$$w_{(x_n,y_n,z_n)}(T,L,F) := \frac{(1-x_n)\,(1-y_n)\,z_n}{1-(1-x_n)\,(1-y_n)\,(1-z_n)}$$

exist as n *tends to infinity. Then*

$$\sum_{\substack{(i_1,i_2,i_3)\in\{(B,L,N),\\(T,R,N),(T,L,F)\}}} \left[\lim_{n\to\infty} w_{(x_n,y_n,z_n)}(i_1,i_2,i_3)\right] = 1.$$

Moreover, as n *tends to infinity,* $\gamma(x_n, y_n, z_n)$ *converges to*

$$\sum_{\substack{(i_1,i_2,i_3)\in\{(B,L,N),\\(T,R,N),(T,L,F)\}}} \left[\lim_{n\to\infty} w_{(x_n,y_n,z_n)}(i_1,i_2,i_3)\right] r(i_1,i_2,i_3).$$

Proof. Part (a) follows from lemma 2.7.2, so it remains to verify part (b). Notice that $w_{(x_n,y_n,z_n)}(B,L,N)$ expresses the probability that, with respect to (x_n, y_n, z_n), the absorption occurs in entry (B, L, N). In fact, the expression $w_{(x_n,y_n,z_n)}(B,L,N)$ equals entry $q_1(t|\,x_n, y_n, z_n)$ of the matrix $Q(x_n, y_n, z_n)$, where t is the absorbing state that entry (B, L, N) leads to. Similar interpretations hold for the other two expressions $w_{(x_n,y_n,z_n)}(T,R,N)$, $w_{(x_n,y_n,z_n)}(T,L,F)$ as well. It is clear that, with respect to (x_n, y_n, z_n), the probability that the eventual absorption occurs in one of the entries (B, L, N), (T, R, N), and (T, L, F) converges to 1 as n tends to infinity. This implies the first equality in (b). The condition that the above limits exist guarantee that $Q(x_n, y_n, z_n)$ has a limit as well, so the second part of (b) follows from lemma 2.7.1-(a). □

Theorem 9.2.3 *There is no stationary* ε*-equilibrium in* Γ *for small* $\varepsilon > 0$.

Proof. Suppose by way of contradiction that (x_n, y_n, z_n) is a stationary ε_n-equilibrium for all $n \in \mathbb{N}$, where ε_n is some positive decreasing sequence converging to 0. Then by using compactness arguments and by taking subsequences, we may assume without loss of generality that (i) the triple (x_n, y_n, z_n) is absorbing for all $n \in \mathbb{N}$ (as stationary ε-equilibria must be absorbing for small $\varepsilon > 0$ due to positive payoffs in entries (B, L, N), (T, R, N), and (T, L, F)), (ii) the sequence (x_n, y_n, z_n) is convergent in the compact space $[0, 1]^3$, and (iii) the expressions

$$w_{(x_n,y_n,z_n)}(B, L, N), \quad w_{(x_n,y_n,z_n)}(T, R, N), \quad w_{(x_n,y_n,z_n)}(T, L, F)$$

in lemma 9.2.2-(b) have limits in the compact space $[0, 1]$. Let

$$(\tilde{x}, \tilde{y}, \tilde{z}) := \lim_{n\to\infty} (x_n, y_n, z_n).$$

We distinguish two essentially different cases. In both cases we derive a contradiction, so the basic assumption that (x_n, y_n, z_n) is a stationary ε_n-equilibrium for all $n \in \mathbb{N}$ will turn out to be false, and then the proof will be complete.

Case 1: $(\tilde{x}, \tilde{y}, \tilde{z})$ is absorbing, namely $\tilde{x} > 0$ or $\tilde{y} > 0$ or $\tilde{z} > 0$.

Suppose without loss of generality that $\tilde{z} > 0$. Then $(\tilde{x}, \tilde{y}, \tilde{z})$ and $(x, \tilde{y}, \tilde{z})$, for all $x \in X$, are absorbing, hence by lemma 9.2.2-(a)

$$\gamma^1(\tilde{x}, \tilde{y}, \tilde{z}) = \lim_{n\to\infty} \gamma^1(x_n, y_n, z_n)$$

$$\gamma^1(x, \tilde{y}, \tilde{z}) = \lim_{n\to\infty} \gamma^1(x, y_n, z_n) \qquad \forall x \in X.$$

Since (x_n, y_n, z_n) is an ε_n-equilibrium, for any $n \in \mathbb{N}$, we have for all $x \in X$

$$\gamma^1(x, y_n, z_n) \leq \gamma^1(x_n, y_n, z_n) + \varepsilon_n,$$

thus we obtain for all $x \in X$ that

$$\begin{aligned} \gamma^1(\tilde{x}, \tilde{y}, \tilde{z}) &= \lim_{n\to\infty} \gamma^1(x_n, y_n, z_n) \\ &\geq \lim_{n\to\infty} \left(\gamma^1(x, y_n, z_n) - \varepsilon_n\right) \\ &= \gamma^1(x, \tilde{y}, \tilde{z}). \end{aligned}$$

For player 2, we have by using analogous arguments that for all $y \in Y$

$$\gamma^2(\tilde{x}, \tilde{y}, \tilde{z}) \geq \gamma^2(\tilde{x}, y, \tilde{z}).$$

If $\tilde{x} > 0$ or $\tilde{y} > 0$, then we obtain similarly for player 3 that for all $z \in Z$

$$\gamma^3(\tilde{x}, \tilde{y}, \tilde{z}) \geq \gamma^3(\tilde{x}, \tilde{y}, z);$$

otherwise for all $z \in Z$

$$\gamma^3(\tilde{x}, \tilde{y}, \tilde{z}) = 1 \geq \gamma^3(\tilde{x}, \tilde{y}, z).$$

Hence $(\tilde{x}, \tilde{y}, \tilde{z})$ is a stationary equilibrium, which contradicts lemma 9.2.1.

Case 2: $(\tilde{x}, \tilde{y}, \tilde{z})$ is non-absorbing, namely $\tilde{x} = \tilde{y} = \tilde{z} = 0$.

By taking a subsequence, due to the symmetrical structure of the game, we may further assume without loss of generality that for all $n \in \mathbb{N}$

$$w_{(x_n,y_n,z_n)}(T, L, F) \geq \max_{(i_1,i_2,i_3)\in\{(B,L,N),(T,R,N)\}} w_{(x_n,y_n,z_n)}(i_1, i_2, i_3) \quad (9.1)$$

and $w_{(0,y_n,z_n)}(T, L, F)$ has a limit as n tends to infinity. By lemma 9.2.2-(b), we have

$$\lim_{n\to\infty} w_{(x_n,y_n,z_n)}(T, L, F) \geq \frac{1}{3}.$$

As

$$3a + b < \frac{3a}{1-b} \qquad \forall a \in [\frac{1}{3}, 1],\ \forall b \in (0, 1),$$

if $\lim_{n\to\infty} w_{(x_n,y_n,z_n)}(B, L, N) > 0$ then we obtain

$$3 \cdot \lim_{n\to\infty} w_{(x_n,y_n,z_n)}(T, L, F) + \lim_{n\to\infty} w_{(x_n,y_n,z_n)}(B, L, N) \quad (9.2)$$

$$< 3 \cdot \frac{\lim_{n\to\infty} w_{(x_n,y_n,z_n)}(T, L, F)}{1 - \lim_{n\to\infty} w_{(x_n,y_n,z_n)}(B, L, N)}.$$

We have

$$w_{(x_n,y_n,z_n)}(T, L, F) = \frac{(1-x_n)(1-y_n)z_n}{1-(1-x_n)(1-y_n)(1-z_n)},$$

$$w_{(x_n,y_n,z_n)}(B, L, N) = \frac{x_n(1-y_n)(1-z_n)}{1-(1-x_n)(1-y_n)(1-z_n)},$$

$$w_{(0,y_n,z_n)}(T, L, F) = \frac{(1-y_n)z_n}{1-(1-y_n)(1-z_n)}.$$

We now show that

$$\lim_{n\to\infty} w_{(x_n,y_n,z_n)}(B,L,N) = 0$$

The opposite

$$\lim_{n\to\infty} w_{(x_n,y_n,z_n)}(B,L,N) > 0$$

would imply, using lemma 9.2.2-(b) and (9.2), that

$$\begin{aligned}
\lim_{n\to\infty} \gamma^1(x_n,y_n,z_n) = \\
&= \lim_{n\to\infty} \left[3 \cdot w_{(x_n,y_n,z_n)}(T,L,F) + w_{(x_n,y_n,z_n)}(B,L,N)\right] \\
&< \lim_{n\to\infty} \left[3 \cdot \frac{w_{(x_n,y_n,z_n)}(T,L,F)}{1 - w_{(x_n,y_n,z_n)}(B,L,N)}\right] \\
&= \lim_{n\to\infty} \left[3 \cdot (1-x_n) \cdot w_{(0,y_n,z_n)}(T,L,F)\right] \\
&= \lim_{n\to\infty} \left[3 \cdot w_{(0,y_n,z_n)}(T,L,F)\right] \\
&= \lim_{n\to\infty} \gamma^1(0,y_n,z_n),
\end{aligned}$$

so (x_n, y_n, z_n) would not be an ε_n-equilibrium for large $n \in \mathbb{N}$. Hence,

$$\lim_{n\to\infty} w_{(x_n,y_n,z_n)}(B,L,N) = 0$$

must hold indeed. But then lemma 9.2.2 and (9.1) yield

$$\lim_{n\to\infty} \gamma^2(x_n,y_n,z_n) = \lim_{n\to\infty} w_{(x_n,y_n,z_n)}(T,R,N) < 1 = \lim_{n\to\infty} \gamma^2(x_n,1,z_n),$$

which is in contradiction with the fact that (x_n, y_n, z_n) is an ε_n-equilibrium for sufficiently large $n \in \mathbb{N}$. □

Now we turn to the class of Markov strategies. First we present a Markov equilibrium, which has a cyclic nature.

Theorem 9.2.4 *In the game Γ, let the Markov strategies κ, λ, μ for players 1,2,3 be respectively given by*

$$\kappa = \left(\frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, \dots\right)$$

$$\lambda = \left(0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \dots\right)$$

$$\mu = \left(0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, \dots\right).$$

Then (κ, λ, μ) is a Markov equilibrium in Γ with equilibrium rewards

$$\gamma(\kappa, \lambda, \mu) = (1, 2, 1).$$

Proof. First notice that

$$\begin{aligned}\gamma(\kappa, \lambda, \mu) &= \frac{1}{2} \cdot (1, 3, 0) + \left(\frac{1}{2}\right)^2 \cdot (0, 1, 3) + \left(\frac{1}{2}\right)^3 \cdot (3, 0, 1) \\ &\quad + \left(\frac{1}{2}\right)^4 \cdot (1, 3, 0) + \left(\frac{1}{2}\right)^5 \cdot (0, 1, 3) + \cdots \\ &= (1, 2, 1).\end{aligned}$$

We prove that κ is a best reply of player 1 against (λ, μ). Similar proofs can be given to show that the other two players have no profitable deviations either, so the proof will then be complete.

First we clarify why player 1 must have a cyclic Markov best reply against (λ, μ). We may represent the cyclic strategies λ and μ as stationary strategies y and z on three non-absorbing states that are identical to the non-absorbing state of the original game Γ, except for entries (T, L, N) which make the play visit these three states cyclically. By applying theorem 2.8.2-(b) to this game with three players, player 1 has a stationary best reply x against (y, z), which, due to the representation, corresponds to a cyclic Markov best reply $\tilde{\kappa}$ against (λ, μ) in the original game Γ.

By way of contradiction suppose now that κ is not a best reply against (λ, μ). Then any best reply $\tilde{\kappa} = (\tilde{x}_n)_{n=1}^{\infty}$ of player 1 against (λ, μ) must be profitable, namely $\gamma^1(\tilde{\kappa}, \lambda, \mu) > \gamma^1(\kappa, \lambda, \mu)$. As we discussed above, we may assume that $\tilde{\kappa}$ is cyclic. We will show that we need to have $\tilde{\kappa} = (0, 0, 0, \dots)$. Let $\tilde{\kappa}_l := (\tilde{x}_n)_{n=l}^{\infty}$, and let $\kappa_l := (x_n)_{n=l}^{\infty}$ for all $l \in \mathbb{N}$. Let λ_l, μ_l be defined analogously. We have

$$\gamma^1(\tilde{\kappa}, \lambda, \mu) > \gamma^1(\kappa, \lambda, \mu) = 1.$$

So $\tilde{x}_1 = 0$, which implies

$$\gamma^1(\tilde{\kappa}_2, \lambda_2, \mu_2) = \gamma^1(\tilde{\kappa}, \lambda, \mu) > 1 = \gamma^1(\kappa_2, \lambda_2, \mu_2),$$

therefore $\tilde{x}_2 = 0$. Then by using

$$\gamma^1(\tilde{\kappa}_2, \lambda_2, \mu_2) = \frac{1}{2}\gamma^1(\tilde{\kappa}_3, \lambda_3, \mu_3),$$

we obtain

$$\gamma^1(\tilde{\kappa}_3, \lambda_3, \mu_3) > 2 = \gamma^1(\kappa_3, \lambda_3, \mu_3),$$

so $\tilde{x}_3 = 0$. Now due to the cyclicity of $\tilde{\kappa}$, we must have $\tilde{\kappa} = (0, 0, 0, \ldots)$ indeed. But then

$$\begin{aligned} \gamma^1(\tilde{\kappa}, \lambda, \mu) &= \frac{1}{2} \cdot 0 + \left(\frac{1}{2}\right)^2 \cdot 3 + \left(\frac{1}{2}\right)^3 \cdot 0 + \left(\frac{1}{2}\right)^4 \cdot 3 + \cdots \\ &= 1 \\ &= \gamma^1(\kappa, \lambda, \mu), \end{aligned}$$

which is in contradiction with the assumption that $\tilde{\kappa}$ is a profitable deviation. So the proof is complete. □

Observe that for all $l \in \mathbb{N}$ the strategies $\kappa_l := (x_n)_{n=l}^{\infty}, \lambda_l := (y_n)_{n=l}^{\infty}, \mu_l := (z_n)_{n=l}^{\infty}$ form cyclic Markov equilibria as well, where κ, λ, and μ are defined as in theorem 9.2.4. Also, if for $\alpha \in [0, 1]$ and $n \in \mathbb{N}$ the notation $\alpha(n)$ represents playing the stationary strategy α for n consecutive stages, then the strategies

$$\pi = (\alpha(n), 0(n), 0(n), \alpha(n), 0(n), 0(n), \alpha(n), \ldots)$$

$$\sigma = (0(n), \alpha(n), 0(n), 0(n), \alpha(n), 0(n), 0(n), \ldots)$$

$$\tau = (0(n), 0(n), \alpha(n), 0(n), 0(n), \alpha(n), 0(n), \ldots)$$

form an equilibrium for each n, if $(1-\alpha)^n = 1/2$. The equality $(1-\alpha)^n = 1/2$ assures that, in any period n of stages, the play absorbs with probability $1/2$.

Notice that, in an equilibrium, any stage where all the players choose their first actions with probability 1 may be skipped without loosing the equilibrium property. The following lemma considers equilibria in terms of strategies where, at any stage, at least one of the players plays his second action with a positive probability.

Lemma 9.2.5 *Let (κ, λ, μ) be a Markov equilibrium in Γ with the property that, at any stage, at least one of the players plays his second action with a positive probability. Then, apart from symmetry, the following properties hold:*

(a) *$x_n, y_n, z_n < 1$ and $(\kappa_n, \lambda_n, \mu_n)$ is an equilibrium for all $n \in \mathbb{N}$;*

(b) *there exists an $n \in \mathbb{N}$ for which $x_n = 0$ or $y_n = 0$ or $z_n = 0$;*

(c) *for any $n \in \mathbb{N}$, if $z_n = 0$ then either $x_n = 0$ or $y_n = 0$;*

(d) *if $x_n > 0$, $y_n = z_n = 0$ then either $x_{n+1} > 0$, $y_{n+1} = z_{n+1} = 0$ or $y_{n+1} > 0$, $x_{n+1} = z_{n+1} = 0$;*

(e) *if $x_n > 0$ and $y_n = z_n = 0$ then $\min\{u_n, v_n, w_n\} = 1$;*

(f) *$x_1 = 0$ or $y_1 = 0$ or $z_1 = 0$.*

Proof. Let

$$\kappa = (x_n)_{n=1}^{\infty}, \quad \lambda = (y_n)_{n=1}^{\infty}, \quad \mu_l = (z_n)_{n=1}^{\infty},$$

and for all $l \in \mathbb{N}$ let

$$\kappa_l := (x_n)_{n=l}^{\infty}, \quad \lambda_l := (y_n)_{n=l}^{\infty}, \quad \mu_l := (z_n)_{n=l}^{\infty}$$

$$(u_l, v_l, w_l) := \gamma(\kappa_l, \lambda_l, \mu_l).$$

Proof of (a): $x_n, y_n, z_n < 1$ and $(\kappa_n, \lambda_n, \mu_n)$ is an equilibrium for all $n \in \mathbb{N}$.

▷ Assume that $x_1 = 1$. Then $y_1 = 0$, hence $z_1 = 1$. This is in contradiction with $x_1 = 1$, therefore $x_1 < 1$ must hold. Due to symmetry, we also have $y_1 < 1$ and $z_1 < 1$.
This means that stage 2 is reached with a positive probability, so $(\kappa_2, \lambda_2, \mu_2)$ is an equilibrium. Therefore $x_2, y_2, z_2 < 1$ must hold as well. Now repeating this argument implies the statement.

Proof of (b): There exists an $n \in \mathbb{N}$ for which $x_n = 0$ or $y_n = 0$ or $z_n = 0$.

▷ Suppose by way of contradiction that $0 < x_n, y_n, z_n$ for all $n \in \mathbb{N}$. Then by (a) we have $0 < x_n, y_n, z_n < 1$ for all $n \in \mathbb{N}$. Now for stage 1 we have

$$u_1 = \gamma^1((0, x_2, x_3, \ldots), \lambda, \mu) = \gamma^1((1, x_2, x_3, \ldots), \lambda, \mu),$$

hence

$$u_1 = 3(1 - y_1)z_1 + y_1 z_1 + (1 - y_1)(1 - z_1)u_2 = 1 - z_1.$$

By expressing u_2

$$u_2 = \frac{1 - 4z_1 + 2y_1 z_1}{(1 - y_1)(1 - z_1)}.$$

Similar equations hold concerning the other two players. Due to symmetry we may assume for stage 1 that $u_1 \leq \min\{v_1, w_1\}$. Then by the equations $u_1 = 1 - z_1,\ v_1 = 1 - x_1,\ w_1 = 1 - y_1$ we obtain $z_1 \geq \max\{x_1, y_1\}$. This implies

$$u_2 \leq 1 - \frac{z_1}{1 - z_1},$$

and then

$$u_1 - u_2 \geq \frac{z_1}{1 - z_1} - z_1 > z_1^2.$$

So we have

$$\begin{aligned} \min\{u_2, v_2, w_2\} &\leq u_2 \\ &< u_1 - z_1^2 \\ &= \min\{u_1, v_1, w_1\} - (\max\{x_1, y_1, z_1\})^2 \\ &< \min\{u_1, v_1, w_1\}. \end{aligned}$$

For stage 2 we have $u_2 = 1 - z_2,\ v_2 = 1 - x_2,\ w_2 = 1 - y_2$, which yields

$$\max\{x_2, y_2, z_2\} > \max\{x_1, y_1, z_1\}.$$

Then analogously

$$\begin{aligned} \min\{u_3, v_3, w_3\} &< \min\{u_2, v_2, w_2\} - (\max\{x_2, y_2, z_2\})^2 \\ &< \min\{u_1, v_1, w_1\} - \left((\max\{x_1, y_1, z_1\})^2 + (\max\{x_2, y_2, z_2\})^2\right), \end{aligned}$$

and

$$\max\{x_3, y_3, z_3\} > \max\{x_2, y_2, z_2\}.$$

In view of the assumption that $0 < x_n, y_n, z_n$ for all $n \in \mathbb{N}$, by using these above arguments inductively, we find that the number $\min\{u_n, v_n, w_n\}$ goes below zero as n increases, which is a contradiction.

Proof of (c): For any $n \in \mathbb{N}$, if $z_n = 0$ then either $x_n = 0$ or $y_n = 0$.

▷ Assume by way of contradiction that $0 < x_n, y_n$. By (a) we have $0 < x_n, y_n < 1$. Then

$$u_n = 1,\ u_{n+1} = \frac{u_n}{1 - y_n} > 1$$

$$v_n = 1 - x_n,\ v_{n+1} = \frac{v_n - 3x_n}{1 - x_n} < 1.$$

Since $u_{n+1} > 1$ we obtain $x_{n+1} = 0$. But then $v_{n+1} \geq 1$, contradiction.

Proof of (d): If $x_n > 0,\ y_n = z_n = 0$ then either $x_{n+1} > 0,\ y_{n+1} = z_{n+1} = 0$ or $y_{n+1} > 0,\ x_{n+1} = z_{n+1} = 0$.

▷ Since $1 \leq w_n = (1 - x_n)w_{n+1}$ we have $w_{n+1} > 1$. The second action of any player cannot give more than 1, so $z_{n+1} = 0$, and by (c) either $x_{n+1} = 0$ or $y_{n+1} = 0$.

Proof of (e): If $x_n > 0$ and $y_n = z_n = 0$ then $\min\{u_n, v_n, w_n\} = 1$.

▷ Obviously, we have $u_n = 1$ and $w_n \geq 1$. Suppose $\bar{n}$ is the first stage after stage n with $y_{\bar{n}} > 0$ (there must be such a stage, otherwise by (d) we would obtain $z_n = z_{n+1} = \cdots = 0$, and hence $\gamma^3(\kappa_n, \lambda_n, \mu_n) = 0 < 1 = \gamma^3(\kappa_n, \lambda_n, 1)$ would hold, contradicting (a)).
By (d) we have

$$x_n, \ldots, x_{\bar{n}-1} > 0, \qquad y_n = z_n = \cdots = y_{\bar{n}-1} = z_{\bar{n}-1} = 0$$

$$x_{\bar{n}} = 0, \qquad y_{\bar{n}} > 0, \qquad z_{\bar{n}} = 0.$$

Thus $\gamma^2(\kappa_{\bar{n}}, \lambda_{\bar{n}}, \mu_{\bar{n}}) = 1$ and

$$v_n = 3\,(x_n + (1 - x_n)x_{n+1} + \cdots + (1 - x_n)\cdots(1 - x_{\bar{n}-2})x_{\bar{n}-1})$$

$$+ (1 - x_n)\cdots(1 - x_{\bar{n}-1})\,\gamma^2(\kappa_{\bar{n}}, \lambda_{\bar{n}}, \mu_{\bar{n}}).$$

Hence $v_n > 1$, so the proof is complete.

Proof of (f): $x_1 = 0$ or $y_1 = 0$ or $z_1 = 0$.

▷ In view of (b), there exists a stage n such that $x_n = 0$ or $y_n = 0$ or $z_n = 0$. Assume that stage n is the first stage with this property. Due to symmetry, we may also assume that $z_n = 0$, and therefore by (c) that $y_n = 0$. Hence $x_n > 0$. Now (e) yields

$$\min\{u_n, v_n, w_n\} = 1.$$

Assume by way of contradiction that $n > 1$. Then by (a) we have

$$0 < x_1, y_1, z_1, \ldots, x_{n-1}, y_{n-1}, z_{n-1} < 1, \quad u_1 = 1 - z_1 < 1.$$

Therefore, as in the proof of (b),

$$\min\{u_n, v_n, w_n\} < \min\{u_1, v_1, w_1\} \leq u_1 < 1,$$

which is a contradiction. So the proof is complete. □

The next theorem says that all Markov equilibria are of the same type as presented in theorem 9.2.4.

Theorem 9.2.6 *Let (κ, λ, μ) be a Markov equilibrium in Γ with the property that, at any stage, at least one of the players plays his second action with a positive probability. Then, at each stage exactly one of the players plays his second action with a positive probability, and these players appear cyclically in the order 1,2,3.*

Proof. Due to symmetry and lemma 9.2.5-(f), we may suppose without loss of generality that $z_1 = 0$. Then by lemma 9.2.5-(c) we have $x_1 = 0$ or $y_1 = 0$. Assume $y_1 = 0$, so $x_1 > 0$. By lemma 9.2.5-(d) either $x_2 > 0$, $y_2 = z_2 = 0$ or $y_2 > 0$, $x_2 = z_2 = 0$. Now using lemma 9.2.5-(d) (and symmetry) inductively yields the statement. □

Theorem 9.2.7 *The set of feasible equilibrium rewards for Γ is the triangle Ψ (without its interior) in $\mathbb{R}^3$ with extreme points $(1,1,2)$, $(1,2,1)$, $(2,1,1)$.*

Proof. Let the strategy triple (κ, λ, μ) be a Markov equilibrium for Γ with rewards $(u, v, w) = \gamma(\kappa, \lambda, \mu)$. We show that $(u, v, w) \in \Psi$. As we discussed, we may assume that, at any stage, at least one of the players plays his second action with a positive probability. By theorem 9.2.6 and lemma 9.2.5-(e), we have $\min\{u, v, w\} = 1$. Hence, it suffices to show the equality $u + v + w = 4$. Since (κ, λ, μ) is an equilibrium, absorption has to occur with probability 1. So in view of theorem 9.2.6, eventually absorption occurs in one of the entries (B, L, N), (T, R, N), (T, L, F). The sum of the payoffs in those entries is 4, thus $u + v + w = 4$ holds indeed. Therefore $(u, v, w) \in \Psi$.
Now we show that if $(u, v, w) \in \Psi$ then there exists a Markov equilibrium (κ, λ, μ) with rewards (u, v, w). By symmetry it suffices to find a Markov equilibrium with rewards $(1, 1 + \alpha, 2 - \alpha)$, where $\alpha \in [0, 1]$. Let

$$\kappa = \left(\frac{\alpha}{2}, 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, \ldots\right)$$

$$\lambda = \left(0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, \ldots\right)$$

$$\mu = \left(0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \ldots\right).$$

These are almost the strategies defined in theorem 9.2.4, but the mixed action of player 1 for the first stage is modified. Now $\gamma(\kappa, \lambda, \mu) = (1, 1 + \alpha, 2 - \alpha)$, and it can be verified similarly to the proof of theorem 9.2.4 that (κ, λ, μ) is an equilibrium indeed. □

9.3 Final remarks

Recently, Solan [1998] proved the existence of ε-equilibria, for all $\varepsilon > 0$, in three-person repeated games with absorbing states. He showed that, in each three-person repeated game with absorbing states, at least one of three essentially different types of ε-equilibria must occur. One of these types is exactly the class of cyclic ε-equilibria, with further interesting results concerning them.

Concluding remarks

We wish to make some concluding remarks about alternative evaluations similar to the average reward as well as some alternative solution concepts regarding the stochastic game model. We will do this in a form of a brief discussion.

In the literature of stochastic games, there are several alternative rewards which, just like the average reward (cf. definition 2.5.1), are frequently used for an evaluation of the long-term average payoffs. Some of the most important ones, for player $k \in \{1,2\}$ and with respect to a strategy pair $(\pi,\sigma) \in \Pi \times \Sigma$ and initial state $s \in S$, are the following :

$$\rho_s^{1,k}(\pi,\sigma) := \mathcal{E}_{s\pi\sigma}\left(\liminf_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} R_n^k\right),$$

$$\rho_s^{2,k}(\pi,\sigma) := \limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathcal{E}_{s\pi\sigma}\left(R_n^k\right),$$

$$\rho_s^{3,k}(\pi,\sigma) := \mathcal{E}_{s\pi\sigma}\left(\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} R_n^k\right),$$

where R_n^k denotes the random variable for the payoff of player k at stage n. We now briefly discuss the relationship between the average reward γ and these above mentioned alternative rewards. First of all, it always holds for any player $k \in \{1,2\}$ that for any $s \in S$, $(\pi,\sigma) \in \Pi \times \Sigma$ we have

$$\rho_s^{1,k}(\pi,\sigma) \leq \gamma_s^k(\pi,\sigma) \leq \rho_s^{2,k}(\pi,\sigma) \leq \rho_s^{3,k}(\pi,\sigma). \qquad \text{(I)}$$

Here the first inequality is a simple consequence of Fatou's lemma (cf. Fatou [1906]), the second inequality follows from the fact that the limit inferior is always smaller than or equal to the limit superior for any real sequences, and

finally the last inequality is an implication of Fatou's lemma together with the fact that the equality $\limsup_{n\to\infty} a_n = -\liminf_{n\to\infty}(-a_n)$ holds for any bounded real sequence $(a_n)_{n\in\mathbb{N}}$.
We wish to stress that equalities in (I) do not hold in general. Nevertheless, it is fortunate to know that all these four rewards are equal when both players use stationary strategies. This is due to the fact that stationary strategy pairs induce Markov processes on the set of states, as discussed in section 2.4. In fact, for any of the alternative rewards ρ^m, $m = 1, 2, 3$, the properties in lemmas 2.7.1 and 2.7.2 remain valid just as theorem 2.8.2-(b) on the best replies against a fixed stationary strategy. Moreover, stationary strategies guarantee the same rewards (cf. definition 2.9.1) irrespective of the choice of these alternative rewards (cf. Bewley & Kohlberg [1978]).

Uniform optimality

In this section, we only consider stochastic games in which the sum of the payoffs is always zero, namely

$$r_s^1(i_s, j_s) = -r_s^2(i_s, j_s) \quad \forall i_s \in I_s,\ \forall j_s \in J_s,\ \forall s \in S;$$

or $r^1 = -r^2$ in a brief notation. Let

$$r_s(i_s, j_s) := r_s^1(i_s, j_s) \quad \forall i_s \in I_s,\ \forall j_s \in J_s,\ \forall s \in S.$$

In these games, by payoffs we will always mean player 1's payoffs. On basis of the above condition, we may assume that these payoffs are payed to player 1 by player 2.
Naturally, with respect to any of the rewards ρ^m, $m = 1, 2, 3$, we can speak of zero-sum stochastic games, values, and optimality, as in section 2.9 with respect to the average reward γ. Before turning to these issues, we wish to define the very appealing concept of uniform (ε-)optimality in zero-sum stochastic games.

Definition. *Suppose that, in a stochastic game, $r := r^1 = -r^2$. Let R_n denote the random variable for the payoff at stage n.*
For a strategy $\pi \in \Pi$ and initial state $s \in S$, let $\underline{w}_s(\pi)$ be the supremum of all the real numbers a_s with the properties that

(a) *for all $\delta > 0$ there exists a stage N^δ such that for all $\sigma \in \Sigma$*

$$\mathcal{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N} R_n\right) \geq a_s - \delta \qquad \forall N \geq N^\delta,$$

(b) *for all* $\sigma \in \Sigma$

$$\mathcal{E}_{s\pi\sigma}\left(\liminf_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}R_n\right) \geq a_s.$$

We say that the strategy π *uniformly guarantees reward* a_s *for initial state* s, *if* $\underline{w}_s(\pi) \geq a_s$.
For a strategy $\sigma \in \Sigma$ *and initial state* $s \in S$, *let* $\overline{w}_s(\sigma)$ *be the infimum of all the real numbers* b_s *with the properties that*

(c) *for all* $\delta > 0$ *there exists a stage* N^δ *such that for all* $\pi \in \Pi$

$$\mathcal{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N}R_n\right) \leq b_s + \delta \qquad \forall N \geq N^\delta,$$

(d) *for all* $\pi \in \Pi$

$$\mathcal{E}_{s\pi\sigma}\left(\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}R_n\right) \leq b_s.$$

We say that the strategy σ *uniformly guarantees reward* b_s *for initial state* s, *if* $\overline{w}_s(\sigma) \leq b_s$.

Bewley & Kohlberg [1978] showed that stationary strategies guarantee the same rewards in the uniform sense as for all the rewards γ and ρ^m, $m = 1, 2, 3$. In particular, $\underline{v}_s(x) = \underline{w}_s(x)$ for all $s \in S$, $x \in X$; and similarly for player 2. It can be shown that the players cannot guarantee better rewards uniformly than with respect to the average reward γ, hence we have

$$\underline{w}_s(\pi) \leq v_s \leq \overline{w}_s(\sigma)$$

for all $s \in S$, $\pi \in \Pi$, $\sigma \in \Sigma$.

Definition. *Suppose that, in a stochastic game,* $r := r^1 = -r^2$. *If there exists a real valued vector* $w = (w_s)_{s\in S}$ *such that*

$$w_s = \sup_{\pi\in\Pi}\underline{w}_s(\pi) = \inf_{\sigma\in\Sigma}\overline{w}_s(\sigma) \qquad \forall s \in S,$$

then w *is called the uniform value of the stochastic game.*

Assume that the uniform value w exists. Then, for initial state $s \in S$, a strategy $\pi \in \Pi$ is called uniformly ε-optimal for player 1, where $\varepsilon \geq 0$, if

$$\underline{w}_s(\pi) \geq w_s - \varepsilon.$$

The strategy π is called uniformly ε-optimal, if it is uniformly ε-optimal for all initial states $s \in S$. Uniformly 0-optimal strategies are simply called uniformly optimal. Similar definitions hold for player 2.

Mertens & Neyman [1981] showed that, in every zero-sum stochastic game, the values for the alternative rewards ρ^m, $m = 1, 2, 3$ and the uniform value exist and they are all equal to the average value v (so particularly $v_s = w_s$ for all $s \in S$). Moreover, the players have strategies, for any $\varepsilon > 0$, that are ε-optimal with regard to the rewards γ and ρ^m, $m = 1, 2, 3$, and at the same time uniformly ε-optimal.

In light of the definition and the above discussion, a uniformly ε-optimal strategy π for player 1 is a strategy with the following properties: (i) for any $\delta > 0$, the strategy π is $(\varepsilon + \delta)$-optimal in the finite game up to stage N, on condition that N is sufficiently large, and at the same time (ii) π is ε-optimal in the infinite game. The intuition behind uniformly ε-optimal strategies for player 2 is similar. So, uniformly ε-optimal strategies can be applied whenever the zero-sum stochastic game is to be played sufficiently long (even over infinitely many stages). As mentioned in Mertens & Neyman [1981], for any $\delta > 0$, uniformly ε-optimal strategies are also $(\varepsilon + \delta)$-optimal with respect to the β-discounted reward if the discount factor $\beta \in (0, 1)$ is sufficiently close to 1. So the main motivation for using uniformly ε-optimal strategies is that their structure is independent of the exact duration of the game or of the exact discount factor (on condition that the game is sufficiently long or the discount factor is large enough).

Note that, using (I) and that the values are equal, average ε-optimality implies ε-optimality for ρ^m, $m = 2, 3$, while uniform ε-optimality yields ε-optimality for any of the rewards γ and ρ^m, $m = 1, 2, 3$.

Now we would like to discuss how the results in chapters $3, 4, 5$ extend to guaranteed rewards and (ε-)optimality for the rewards ρ^m, $m = 1, 2, 3$, as well as to uniformly guaranteed rewards and uniform (ε-)optimality. Based on the previous discussion, the extensions to the rewards ρ^m, $m = 2, 3$, are immediate. Hence we only focus on reward ρ^1, uniformly guaranteed rewards, and uniform (ε-)optimality. In the light of the observations above, the extensions of the results concerning stationary strategies are straightforward, so we only need

to examine the results in these chapters where no stationary strategies are involved.
First of all, in the condition of Main Theorem 3 in chapter 3, it makes no difference in which sense player 1 has an optimal strategy. However, the Markov strategy f in the second part of Main Theorem 3 is, unfortunately, not necessarily optimal for ρ^1 nor uniformly optimal. Nevertheless, its construction in the proof of theorem 3.3.1-(b) guarantees that for all $\delta > 0$ there exists a stage N^δ such that for all $\sigma \in \Sigma$

$$\mathcal{E}_{sf\sigma}\left(\frac{1}{N}\sum_{n=1}^{N} R_n\right) \geq v_s - \delta \qquad \forall N \geq N^\delta.$$

In order to achieve optimality for ρ^1, which would now also be sufficient for uniform optimality by the above inequality, a somewhat more subtle but rather technical construction can be given.
Because the results and the techniques in chapter 4 are similar to those in chapter 3, the same can be said about Main Theorem 4.
In chapter 5, for any $\varepsilon > 0$, if $K \in \mathbb{N}$ is sufficiently large then the Markov strategy f^K constructed for theorem 5.2.2 is also ε-optimal for reward ρ^1 (see inequalities (5.13) in the proof of lemma 5.2.11). Although the proofs suggest that f^K should be uniformly ε-optimal as well, it does not immediately follow from the proven results.

Uniform equilibria

With respect to the rewards ρ^m, $m = 1, 2, 3$, we can investigate general-sum stochastic games and we may define (ε-)equilibria, as in section 2.10 for the average reward γ. But first we define the concept of uniform (ε-)equilibria.

Definition. *In a general-sum stochastic game, for initial state $s \in S$, a pair of strategies $(\pi, \sigma) \in \Pi \times \Sigma$ is called a uniform ε-equilibrium, $\varepsilon \geq 0$, with reward $a = (a^1, a^2)$, if for all $\delta > 0$ there exists a stage N^δ with the following properties*

(a) *for both players $k = 1, 2$*

$$a_s^k - \delta \leq \mathcal{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N} R_n^k\right) \leq a_s^k + \delta \qquad \forall N \geq N^\delta,$$

$$\mathcal{E}_{s\pi\sigma}\left(\liminf_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}R_n^k\right)=\mathcal{E}_{s\pi\sigma}\left(\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}R_n^k\right)=a_s^k;$$

(b) *for all* $\bar{\pi}\in\Pi$

$$\mathcal{E}_{s\bar{\pi}\sigma}\left(\frac{1}{N}\sum_{n=1}^{N}R_n^1\right)\le\mathcal{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N}R_n^1\right)+\varepsilon+\delta\qquad\forall N\ge N^\delta,$$

$$\mathcal{E}_{s\bar{\pi}\sigma}\left(\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}R_n^1\right)\le\mathcal{E}_{s\pi\sigma}\left(\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}R_n^1\right)+\varepsilon;$$

(c) *for all* $\bar{\sigma}\in\Sigma$

$$\mathcal{E}_{s\pi\bar{\sigma}}\left(\frac{1}{N}\sum_{n=1}^{N}R_n^2\right)\le\mathcal{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N}R_n^2\right)+\varepsilon+\delta\qquad\forall N\ge N^\delta,$$

$$\mathcal{E}_{s\pi\bar{\sigma}}\left(\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}R_n^2\right)\le\mathcal{E}_{s\pi\sigma}\left(\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}R_n^2\right)+\varepsilon,$$

where R_n^k *denotes the random variable for the payoff for player* k *at stage* n. *The strategy pair* (π,σ) *is a uniform* ε*-equilibrium, if it is a uniform* ε*-equilibrium for all initial states* $s\in S$. *Uniform* 0*-equilibria are simply called uniform equilibria.*

A uniform ε-equilibrium (π,σ) is thus a pair of strategies with the following properties: (i) for any $\delta>0$, the pair (π,σ) forms an $(\varepsilon+\delta)$-equilibrium in the finite game up to stage N, on condition that N is sufficiently large, and at the same time (ii) (π,σ) is an ε-equilibrium in the infinite game. The motivation for applying uniform ε-equilibria is quite the same as for uniformly ε-optimal strategies. Uniform ε-equilibria can therefore be used whenever the stochastic game is to be played sufficiently long (even over infinitely many stages). Moreover, for any $\delta>0$, uniform ε-equilibria are also $(\varepsilon+\delta)$-equilibria with respect to the discounted rewards on condition that the discount factors are sufficiently close to 1. So their appealing property is that their structure and the corresponding rewards (up to some small δ) are independent of the

exact duration of the game or of the exact discount factors (on condition that the game is sufficiently long or the discount factors are large enough).
Note that, using (I), uniform (ε-)equilibria are necessarily (ε-)equilibria for any of the rewards γ and ρ^m, $m = 1, 2, 3$.
We will now briefly discuss how the results in chapters $6, 7, 8, 9$ extend to uniform (ε-)equilibria, and therefore to (ε-)equilibria for the rewards ρ^m, $m = 1, 2, 3$. Recall that the rewards γ and ρ^m, $m = 1, 2, 3$, coincide when stationary strategies are used and that stationary strategies guarantee the same rewards for γ and ρ^m, $m = 1, 2, 3$ as well as in the uniform sense.
On basis of the above observations, one can show the validity of Main Theorem 6 for almost stationary uniform ε-equilibria. By almost stationary uniform ε-equilibria we obviously mean uniform ε-equilibria which have the almost stationary property specified in definition 6.1.1. Note that the existence of uniformly ε-optimal strategies in zero-sum games plays a crucial role for this extension here.
In chapter 7, only stationary strategies are used, so the results are valid in the uniform sense as well.
As for chapter 8, we may define average-discounted uniform (ε-)equilibria as average-discounted (ε-)equilibria (cf. definition 8.1.1) with an additional uniformity property on the side of player 1 (recall that player 1 uses the average reward, while player 2 is interested in his discounted reward). It can be verified that Main Theorem 8 generalizes to the uniform case, based on the structure of the strategies constructed there.
In chapter 9, the results of Main Theorem 9 hold in the uniform sense as well. It is important that all Markov equilibria in the game in example 9.1.1 must eventually lead to absorption with probability 1.

References

Bewley T & Kohlberg E [1976,I]: The asymptotic theory of stochastic games. *Mathematics of Operations Research* 1, 197-208.

Bewley T & Kohlberg E [1976,II]: The asymptotic solution of a recursive equation occurring in stochastic games. *Mathematics of Operations Research* 1, 321-336.

Bewley T & Kohlberg E [1978]: On stochastic games with stationary optimal strategies. *Mathematics of Operations Research* 3, 104-125.

Blackwell D [1962]: Discrete dynamic programming. *Annals of Mathematical Statistics* 33, 719-726.

Blackwell D & Ferguson TS [1968]: The big match. *Annals of Mathematical Statistics* 39, 159-163.

Bohnenblust HF, Karlin S & Shapley LS [1950]: Solutions of discrete two-person games. In: Kuhn HW & Tucker AW (eds.), Contributions to the theory of games I, *Annals of Mathematical Studies* 24, Princeton University Press, 51-72.

Coulomb JM [1992]: Repeated games with absorbing states and no signals. *International Journal of Game Theory* 21, 161-174.

Doob JL [1953]: *Stochastic processes.* Wiley, New York.

Dudley RM [1989]: *Real analysis and probability.* Wadsworth, Inc., Belmont, California.

Everett H [1957]: Recursive games. In: Dresher M, Tucker AW & Wolfe P (eds.), Contributions to the theory of games III, *Annals of Mathematical Studies* 39, Princeton University Press, 47-78.

Fatou PJL [1906]: Séries trigonométriques et séries de Taylor. *Acta Mathematica* 30, 335-400.

Federgruen A [1978]: On n-person stochastic games with denumerale state space. *Advances in Applied Probability* 10, 452-471.

Feinberg EA & Shwartz A [1994]: Markov decision models with weighted discounted criteria. *Mathematics of Operations Research* 19, 152-168.

Feinberg EA & Shwartz A [1995]: Constrained Markov decision models with weighted discounted rewards. *Mathematics of Operations Research* 20, 302-320.

Filar JA [1981]: Ordered field property for stochastic games when the player who controls transitions changes from state to state. *Journal of Optimization Theory and Applications* 34, 503-515.

Filar JA & Vrieze OJ [1992]: Weighted reward criteria in competitive Markov decision processes. *Zeitschrift für Operations Research* - Methods and models of operations research 36, 343-358.

Fink AM [1964]: Equilibrium in a stochastic n-person game. *Journal of Science of Hiroshima University,* Series A-I 28, 89-93.

Flesch J & Perea y Monsuwé A [1997]: Repeated games with endogenous choice of information mechanisms. Report, Maastricht University.

Flesch J, Thuijsman F & Vrieze OJ [1996]: Recursive repeated games with absorbing states. *Mathematics of Operations Research* 21, 1016-1022.

Flesch J, Thuijsman F & Vrieze OJ [1997,I]: Cyclic Markov equilibria in a cubic game. *International Journal of Game Theory* 26, 303-314.

Flesch J, Thuijsman F & Vrieze OJ [1997,II]: Markov strategies are better than stationary strategies. *International Game Theory Review*, to appear.

Flesch J, Thuijsman F & Vrieze OJ [1998,I]:. Simplifying optimal strategies in stochastic games. *SIAM Journal on Control and Optimization* 36, No. 4, 1331-1347.

Flesch J, Thuijsman F & Vrieze OJ [1998,II]: Almost stationary ε-equilibria in zero-sum stochastic games. *Journal of Optimization Theory and Applications,* to appear.

Flesch J, Thuijsman F & Vrieze OJ [1998,III]: Average-discounted equilibria in stochastic games. *European Journal of Operational Research,* to appear.

Flesch J, Thuijsman F & Vrieze OJ [1998,IV]: Improving strategies in stochastic games. *Proceedings of the 37th Conference on Decision and Control,* to appear.

Gale D & Sherman S [1950]: Solutions of finite two-person games. In: Kuhn HW & Tucker AW (eds.), Contributions to the theory of games I, *Annals of Mathematical Studies* 24, Princeton University Press, 37-49.

Gillette D [1957]: Stochastic games with zero stop probabilities. In: Dresher M, Tucker AW & Wolfe P (eds.), Contributions to the theory of games III, *Annals of Mathematical Studies* 39, Princeton University Press, 179-187.

Hoffman AJ & Karp RM [1966]: On nonterminating stochastic games. *Management Science* 12, 359-370.

Hordijk A, Kallenberg LCM & Wanrooij GL [1983]: Semi-Markov strategies in stochastic games. *International Journal of Game Theory* 12, 81-89.

Kakutani S [1941]: A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* 8, 416-427.

Kemeny J & Snell J [1960]: *Finite Markov chains.* Van Nostrand, Princeton.

Kohlberg E [1974]: Repeated games with absorbing states. *Annals of Statistics* 2, 724-738.

Kolmogorov A [1933]: Grundbegriffe der wahrscheinlichkeitsrechnung. *Ergebnisse der Mathematik 2,* No. 3, Springer Verlag, Berlin.

Liggett TM & Lippman SA [1969]: Stochastic games with perfect information and time average payoff. *SIAM Review* 11, 604-607.

Myerson RB [1978]: Refinements of the Nash equilibrium concept. *International Journal of Game Theory* 7, 73-80.

Mertens JF & Neyman A [1981]: Stochastic games. *International Journal of Game Theory* 10, 53-66.

Monash CA [1980]: *Stochastic games: the minimax theorem.* Ph.D. thesis, Harvard University, Cambridge, Massachusetts.

Nowak AS & Raghavan TES [1991]: Positive stochastic games and a theorem of Ornstein. In: Raghavan TES, Ferguson TS, Vrieze OJ & Parthasarathy T (eds.), *Stochastic Games and Related Topics*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 127-134.

Parthasarathy T, Tijs SH & Vrieze OJ [1984]: Stochastic games with state independent transitions and separable rewards. In: Hammer G & Pallaschke D (eds.), *Selected Topics in Operations Research and Mathematical Economics*, Springer Verlag, Berlin 262-271.

Raghavan TES, Tijs SH & Vrieze OJ [1985]: On stochastic games with additive reward and transition structure. *Journal of Optimization Theory and Applications* 47, 451-464.

Rogers PD [1969]: *Non-zerosum stochastic games.* Ph.D. thesis, Report ORC 69-8, Operations Research Center, University of California, Berkeley.

Shapley LS [1953]: Stochastic games. *Proceedings of the National Academy of Sciences U.S.A.* 39, 1095-1100.

Sobel MJ [1971]: Noncooperative stochastic games. *Annals of Mathematical Statistics* 42, 1930-1935.

Solan E [1998]: *Stochastic games.* Ph.D. thesis, Hebrew University, Jerusalem.

Sorin S [1986]: Asymptotic properties of a non-zerosum game. *International Journal of Game Theory,* 15, 101-107.

Schweitzer PJ [1968]: Perturbation theory and finite Markov chains. *Journal of Applied Probability,* 5, 401-41.

Takahashi M [1964]: Equilibrium points of stochastic noncooperative n-person games. *Journal of Science of Hiroshima University,* Series A-I 28, 95-99.

Thuijsman F [1992]: *Optimality and equilibria in stochastic games.* CWI-Tract 82, Centre for Mathematics and Computer Science, Amsterdam.

Thuijsman F & Raghavan TES [1997]: Perfect information stochastic games and related classes. *International Journal of Game Theory,* 26, 403-408.

Thuijsman F & Vrieze OJ [1992]: Note on recursive games. In: Dutta B, Mookherjee D, Parthasarathy T, Raghavan TES, Ray D & Tijs SH (eds.), *Game Theory and Economic Applications*, Lecture Notes in Economics and Mathematical Systems, Springer Verlag, Berlin, 389, 133-145.

Thuijsman F & Vrieze OJ [1993]: Stationary ε-optimal strategies in stochastic games., *OR Spektrum*, Springer Verlag, 15, 9-15.

Thuijsman F & Vrieze OJ [1998]: The power of threats in stochastic games. In: Bardi M, Raghavan TES & Parthasarathy T (eds.), *Stochastic and Differential Games, Theory and Numerical Methods*, Birkhauser, Boston.

van Damme E [1991]: *Stability and perfection of Nash equilibria.* Springer Verlag, Berlin.

Vieille N [1993]: Solvable states in stochastic games. *International Journal of Game Theory* 21, 395-404.

Vieille N [1994]: On equilibria in undiscounted stochastic games. D.P. 9446, Ceremade.

Vieille N [1997,I]: 2-person stochastic games I: A reduction, D.P. 9745, Ceremade.

Vieille N [1997,II]: 2-person stochastic games II: The case of recursive games, D.P. 9747, Ceremade.

von Neumann J [1928]: Zur theorie der gesellschaftsspiele. *Mathematische Annalen* 100, 295-320.

Vrieze OJ & Thuijsman F [1989]: On equilibria in repeated games with absorbing states. *International Journal of Game Theory* 18, 293-310.

Index

Bewley T, 24, 50, 152, 153
Blackwell D, 21, 22, 24, 26
Bohnenblust HF, 39

Coulomb JM, 70, 88

Doob JL, 11
Dudley RM, 10, 79, 81

Everett H, 32

Fatou PJL, 87, 151
Federgruen A, 33
Feinberg EA, 134
Ferguson TS, 24, 26
Filar JA, 32, 134
Fink AM, 27, 28
Flesch J, 37, 58, 70, 96, 111, 123, 135

Gale D, 39
Gillette D, 12, 24

Hoffman AJ, 32
Hordijk A, 21, 22

Kakutani S, 116, 122
Kallenberg LCM, 21, 22
Karlin S, 39
Karp RM, 32
Kemeny J, 11
Kohlberg E, 24, 32, 50, 88, 152, 153
Kolmogorov A, 10

Liggett TM, 32, 33
Lippman SA, 32, 33

Mertens JF, 24, 40, 154
Monash CA, 21
Myerson RB, 115

Neyman A, 24, 40, 154
Nowak AS, 94

Parthasarathy T, 32, 33

Raghavan TES, 32, 33, 94, 133
Rogers PD, 33

Schweitzer PJ, 15
Shapley LS, 13, 22, 24, 39
Sherman S, 39
Shwartz A, 134
Snell J, 11
Sobel MJ, 33
Solan E, 149
Sorin S, 30

Takahashi M, 27, 28
Thuijsman F, 18, 32, 33, 37, 55, 58, 70, 88, 92, 95, 96, 99, 103, 111, 123, 133–135
Tijs SH, 32, 33

van Damme E, 115
Vieille N, 30, 33, 95, 103
von Neumann J, 38

Vrieze OJ, 32, 33, 37, 55, 58, 70, 88, 92, 95, 96, 99, 103, 111, 123, 134, 135

Wanrooij GL, 21, 22

Samenvatting

Dit proefschrift levert nieuwe theoretische inzichten op het gebied van stochastische spelen. Een stochastisch spel kan gezien worden als een beslissingsproces, waarin de deelnemers (ook wel spelers genoemd) de beslissingen maken. Wanneer er maar één speler is, dan kan diegene haar beslissingen zonder concurrentiestrijd nemen. Dit is een speciaal probleem dat in de literatuur bekend staat als Markov beslissings problemen. In het verdere verloop van deze samenvatting zullen we altijd aannemen dat er minstens twee spelers zijn, die elkaars resultaten kunnen beïnvloeden.
We gaan nu het model van stochatische spelen met twee spelers beschrijven. Het model kan ook makkelijk uitgebreid worden voor meerdere spelers. Een twee-persoons stochastisch spel begint in een bepaalde initiële positite. In deze positie moet elke speler een zet kiezen uit haar verzameling van mogelijke zetten. Vereiste is, dat de zet van de andere speler pas na het maken van de eigen beslissing bekend wordt. Afhankelijk van de zetten van beide spelers, krijgt elke speler bepaalde inkomsten (indien het bedrag negatief is, moet het als uitgave worden gezien) en wordt er een kans toegekend aan alle mogelijke vervolg posities. Met behulp van deze kansen wordt een volgende positie toegewezen. In deze tweede positie moeten de spelers weer zetten kiezen. En net zoals in de initiële positie zullen de spelers op basis van de keuzes bepaalde inkomsten krijgen, waarna het spel weer naar een volgende positie gaat. Dit gaat zo door tot een bepaald aantal zetten gedaan is (wat eventueel ook oneindig kan zijn).
Het volgende economische voorbeeld zal het model van stochastische spelen illustreren. Veronderstel dat er twee ondernemingen op eenzelfde markt actief zijn, en dat ze beide hun eigen wekelijkse plannen maken. Dan zullen de inkomsten van de ondernemingen natuurlijk afhankelijk zijn van beide plannen. Verder zullen de ondernemingen de positie van de markt-situatie beïnvloeden door hun plannen uit te voeren.

Het bovenstaande theoretische model is dus toepasbaar, wanneer we de ondernemingen gelijk stellen aan spelers, de markt-situatie aan de positie en het uitvoeren van de wekelijkse plannen zien als een zet.
Elke speler heeft als doel om zijn eigen gemiddelde inkomsten te maximaliseren. Vooruitkijken is cruciaal voor de spelers, want korststondig succes is geen garantie voor een goede toekomst (hardlopers zijn doodlopers). Dus op basis van het spel-verloop moet de speler erop letten dat zowel haar inkomst als de komende positie gunstig is. Aangezien de belangen van de spelers niet altijd overeenkomstig zijn, kunnen er zo conflict situaties optreden.
Stochastische spelen zijn niet-coöperatieve spelen in de zin dat de spelers niet mogen samenwerken om hoge gemiddelde inkomsten te verkrijgen. Elke speler zal daarom met haar eigen strategie haar zetten moeten bepalen. De theorie is bezig met het zoeken naar en analyseren van strategie-paren die de eigenschap hebben dat geen speler haar eigen gemiddelde inkomsten kan verbeteren door een andere stategie te kiezen. Deze paren worden evenwichtig genoemd op grond van de bovenstaande eigenschap.
Het is theoretisch nog niet bewezen dat evenwichten altijd bestaan, maar hun existentie is wel bekend in een aantal speciale klassen van spelen. Dit proefschrift verrijkt de theorie met verdere resultaten omtrent de existentie en de structuur van evenwichten.

Összefoglalás

Ezen doktori disszertáció a sztochasztikus játékok elméletébe nyújt betekintést. Egy sztochasztikus játék egy döntési folyamatnak tekinthető, amelyben a döntéseket a benne résztvevő játékosok hozzák meg. Amennyiben csak egy résztvevő játékos van, akkor ő érdekellentét nélkül hozhatja meg döntéseit, s így a probléma egy sajátos jelleget ölt (a teljesség kedvéért megjegyezzük, hogy az egyszemélyes sztochasztikus játékok mint Markov döntési folyamatok ismertek az irodalomban). A továbbiakban ezért mindig feltételezni fogjuk, hogy a játékosok száma legalább kettő.

Az alábbiakban a kétszemélyes sztochasztikus játékok modelljét fogjuk csak ismertetni, mivel a modell könnyen kiterjeszthető több játékos esetére is. Egy kétszemélyes sztochasztikus játék egy meghatározott kezdeti pozícióból indul. Ebben a pozícióban mindkét játékosnak egy lépést kell kiválasztania a számára lehetséges lépések halmazából, azzal a megkötéssel, hogy a másik lépését csak saját döntésének meghozatala után ismeri meg. A megtett lépésektől függően a játékosok ezt követően bevételhez jutnak - amennyiben ez az összeg negatív szám, akkor az természetesen kiadásnak tekintendő. Ezután, úgyszintén a lépések függvényében, minden egyes lehetséges pozícióhoz egy kiválasztási valószínűség rendelődik, s ez alapján a játék egy következő pozícióba kerül. Ebben a második pozícióban a két játékosnak ismét lépnie kell, s csakúgy mint a kezdeti pozícióban, ezen lépésektől függően a játékosok bevételhez jutnak, majd a játék egy harmadik pozícióba jut. A játék a fent leírt módon folytatódik tovább egy meghatározott számú lépésig vagy a végtelenségig.

A következő közgazdaságtani példa jól illusztrálja a sztochasztikus játékok modelljét. Tegyük fel, hogy két vállalat ugyanazon a piacon érdekelt, és minden egyes hétre egy új gazdasági tervet készít. A vállalatok bevétele természetesen az általuk választott gazdasági tervektől függ. A vállalatok a gazdasági terveiken keresztül persze a piaci helyzetet is befolyásolni tudják.

A párhuzam a példa es az elmélet között a következő: a vállalatokat mint a játékosokat, a piaci helyzetet mint a pozíciót, míg a gazdasági terveket mint a játékosok lépéseit képzelhetjük el.
Mindkét játékos arra törekszik, hogy az ő saját átlagos bevétele minél nagyobb legyen. A játék lefolyása ismerétében világos tehát, hogy egy lépés kiválasztásakor nem csak az akkori bevételét kell szem előtt tartania, hanem egyben arra is ügyelnie kell, hogy a játék számára kedvező pozícióba jusson. Lévén, hogy a játékosok érdekei nem feltétlenül egyeznek meg, a játék alatt érdekellentétek léphetnek fel.
A sztochasztikus játékok nem kooperatív játékok, ami azt jelenti, hogy a játékosok nem szövetkezhetnek annak érdekében, hogy magas átlagos bevételeket biztosítsanak maguknak. Más szavakkal, mindkét játékosnak a saját lépéseinek a kiválasztásához egy önálló stratégiát kell alkalmaznia. Az elmélet tehát olyan stratégiapárok keresésével és elemzésével foglalkozik, amelyeket az jellemez, hogy egyik játékos sem tudná a saját átlagos bevételét növelni ha egyedül egy másik stratégiára áttérne. Ezeket a párokat az iménti tulajdonságuk alapján egyensúlyi stratégiapároknak nevezzük.
Az egyensúlyi stratégiapárok létezése még elméletileg mindig nem bizonyított, de már bizonyos típusú játékok esetében ismert. Ezen doktori disszertáció célja további eredmények feltárása az egyensúlyi stratégiapárok létezésének és szerkezetének irányában.

About the author

János Flesch was born on July 19th, 1970 in Budapest, Hungary. After attending the Tóth Árpád secondary school from 1984 till 1988, he went on to study mathematics at the Kossuth Lajos University in Debrecen. In his final year, he participated in the Master Class courses on functional analysis, organized by the Mathematical Research Institute in the Netherlands. Upon successful completion of his Master's thesis on chess programming (in artificial intelligence) under the supervision of Dr. Magda Várterész, in July 1994, he started working on game theory as a research assistant at the Department of Mathematics, Maastricht University.